Le traitement automatique des langues dans les industries de l'information

Livre blanc

Groupe de Réflexion sur
les Industries de l'Information et
les Industries de la Langue
(GRIIIL)

Janvier 2005

Avant-propos

Alain Couillault

Ce livre blanc a été réalisé dans le cadre du Groupe de Réflexion sur les Industries de l'Information et les Industries de la Langue, organisé conjointement par l'Association des Professionnels des Industries de la Langue (APIL) et le Groupement Français de l'Industrie de l'Information (GFII).

Il s'adresse tout d'abord aux utilisateurs, présents ou futurs, d'applications ou d'outils utilisant le traitement automatique des langues, et en présente les usages possibles et les bénéfices qu'ils peuvent en attendre.

Il s'adresse également aux professionnels des industries de la langue et à ceux des Industries de l'Information, auxquels il propose une vue synthétique de l'usage de l'information dans et hors l'entreprise.

Ce livre est le fruit des réflexions et témoignages apportés par plusieurs acteurs, individus ou entreprises, des deux domaines concernés.

Alain Couillault et Alain Garnier ont animé ce groupe de travail.

Ils ont participé à sa rédaction, ainsi que Eric Debonne, Gil Francopoulo, Fabienne Gire, Sylvie Guillemin Lanne, Claude de Loupy, Guillaume Mazières, Bruno Menon et Hugues Sézille de Mazancourt.

Les participants au Groupe de réflexion comprennent également Alexandre Arcouteil, Philippe Bonny, Stéphane Chaudiron, Carole Chevalier, Khalid Choukri, Martine Dejean, Jean Delahousse, Luc Grivel, Isabelle Leclercq, Laurent Lefoll, Bernard Normier, Ruth Martinez, Christine Reynaud, Lionel Stouder, Laurence Zaysser.

Véronique Zablot a rassemblé les différentes contributions et mis en forme ce document.

Introduction

Alain Garnier

Le propre de l'homme, au delà du rire, est bien d'avoir une capacité intellectuelle qui se déploie depuis des millénaires via les langues dites « naturelles ». Le langage se retrouve aujourd'hui au cœur de tous les systèmes d'information : sous forme orale quand il s'agit du téléphone, ou écrite pour le mail, les fichiers ou encore le Web. Or la compétition sans cesse croissante des entreprises et des organisations pose la question de l'usage du « langage naturel » au sein des systèmes d'information comme une source de productivité potentielle. Aussi, la diversité à la fois de forme mais aussi de fond donne lieu à un challenge fantastique : faire en sorte que la faculté de langage soit prolongée au sein des machines. C'est un défi technologique qui vise à apporter des réponses concrètes à la gestion sans cesse croissante de textes, documents, mails, productions sonores et vidéo. Comment fournir cette convergence concernant le langage que chacun attend à travers des besoins aussi divers que : rechercher, classer, analyser, diffuser, reproduire, vérifier… ?

Le Traitement automatique des Langues (TAL) fédère un ensemble d'acteurs visant à mutualiser les avancées et les capacités des systèmes d'information, dont l'objet est d'apporter un service pour le traitement de l'information non structurée. Or, l'information issue du langage étant présente dans de nombreux systèmes d'information, sans exclusive quant au secteur d'activité c'est fort logiquement que les technologies embarquées sont utilisées dans de nombreux secteurs de l'activité des entreprises. Ces technologies sont également présentes, de manière grandissante, et sans être toujours évidentes pour l'utilisateur, dans divers systèmes grands public.

Ce document a pour vocation de fournir aux décideurs qui sont confrontés à une problématique liée à la gestion de l'information des clés de compréhension pour établir en quoi et pour quels usages les technologies du traitement automatique des langues sont utiles et pertinentes.

Le présent document est décomposé en trois grandes parties. La première établit un panorama de l'usage du traitement automatique des langues aujourd'hui, que ce soit dans les entreprises, dans les services embarqués ou pour les professionnels du domaine.

La seconde partie déroule des applications concrètes et instanciées des technologies au sein des entreprises, en donnant, pour chacune d'elles, un exemple concret et une métrique permettant d'évaluer le bien fondé de l'utilisation du TAL dans un projet de ce type. C'est une grille de lecture à la fois didactique et discriminante qui permet à un chef de projet, décideur ou consultant, de mesurer l'impact du traitement automatique des langues, dans des contextes variés.

Enfin, la troisième partie traite des standards qui émergent autour du TAL, notamment le Web Sémantique, afin de coordonner les efforts des industriels mais également des utilisateurs et des chercheurs.

Ce document se veut d'un usage pratique et pragmatique dans un contexte opérationnel. Le TAL qui fête bientôt ses trente ans a l'âge de raison qui permet un tel tour d'horizon sur le chemin accompli et celui qui s'ouvre à lui.

Panorama du marché industriel du Traitement Automatique des Langues

Pour fournir un panorama de l'usage et de l'utilisation du traitement automatique des langues, il est nécessaire d'appréhender de manière transverse le « marché » global des technologies de l'information.

Afin de structurer cette démarche, nous proposons une grille de lecture en trois axes :

Le premier axe concerne l'usage *in-situ* du TAL au sein des acteurs économiques : entreprises ou assimilées (ministères, organismes…).

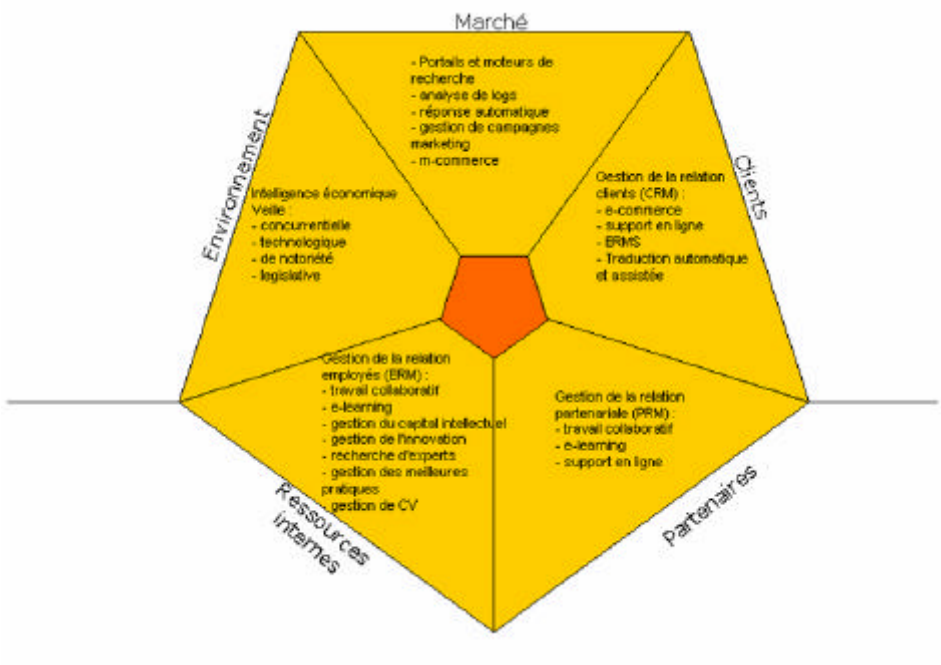
Le deuxième axe concerne le TAL embarqué au cœur des produits, appareils grand public qui nous entourent (téléphones, voitures…).

Enfin le troisième axe concerne les professionnels de l'Information (éditeurs, agences de presse, médias …) dont le métier est l'information et qui ont donc à traiter systématiquement du langage naturel dans leur chaîne de valorisation de l'Information.

Le TAL dans l'entreprise

Alain Couillault

Une entreprise est un « corps sociologique » complexe, aux interactions multiples. Le schéma ci-dessous tente de représenter les interfaces qu'une entreprise entretient avec son environnement, chaque interface étant un lieu de circulation d'information, potentiellement non structurée. A ce titre, ces interfaces constituent des applications des Industries de la Langue, et concernent les Industries de l'Information.



Le centre de ce schéma représente l'entreprise, l'extérieur représente son environnement. Les différents secteurs du schéma listent ainsi différentes activités de l'entreprise :

- en bas, celles qui sont liées à ses moyens de production, qu'ils soient propres ("ressources internes") ou externes ("partenaires") ;
- en haut, celles qui sont liées à son marché : son environnement concurrentiel et technologique ("environnement"), ses prospects ("marché") ou ses clients.

Les ressources internes

Les ressources internes sont aujourd hui une « matière première » pour les entreprises et constituent une part importante de leur « capital intangible ». Or, ce capital est composé à 80% de données dites « non structurées ». Aussi les processus de gestion, d optimisation, de création de ces éléments qui forment les ressources internes, utilisent-ils systématiquement, de manière plus ou moins directe, les technologies de la langue. Ces ressources internes peuvent être liées à des moyens de productions spécifiques (processus industriels, centres de services...), pour lesquels les industries de la langue peuvent avoir des applications spécifiques. Les applications sont généralement liées à la gestion de la relation avec les employés. Cela comprend (d'une certaine façon, dans l'ordre des étapes de la relation avec un employé) la recherche et la gestion des curriculum vitae, la formation en ligne (e-learning), la gestion électronique de documents (GED), le travail collaboratif, la gestion de l'innovation ou la collecte des meilleures pratiques.

Les partenaires

La gestion des partenaires est semblable, en quelque sorte, à la gestion des ressources internes et à la relation clients : les partenaires sont à la fois des producteurs au service de l'entreprise et, tout en y étant externes, entretiennent une relation contractuelle avec celle-ci. Ainsi, les applications des technologies du traitement automatique des langues comprennent aussi bien la formation en ligne et le travail collaboratif, que le support en ligne.

Les clients

La gestion de la relation clients (CRM) comprend différents aspects dans lesquels les industries de la langue peuvent être impliquées : e-commerce, support en ligne, traduction automatique pour ne citer que les plus emblématiques. On a choisi, par exemple, d'inclure ici les traductions automatiques et assistées, qui permettent de fournir de la documentation à une clientèle multilingue. Les systèmes d'aide au support comportent différentes fonctions pour lesquelles le traitement de la langue écrite ou orale apportent de réels gains, ou sont nécessaires : automates de reconnaissance et synthèse vocale pour le support automatique par téléphone, le traitement automatique des courriers électroniques (ERMS), ou la constitution et l'interrogation de bases de connaissances qui sont autant d'exemples de ce type d'applications.

Les clients potentiels

Les technologies du traitement automatique des langues sont utiles à une meilleure relation avec chacun des individus (et les applications se rapprochent alors de la gestion de la relation client) ou avec des nombres importants de prospects (il s'agit alors de marketing) car elles permettent de « parler le langage du client » dans les échanges électroniques qui sont envisagés. Dans la première catégorie se trouvent les applications

liées, globalement, aux moteurs de recherche, qu'elles concernent l'aide à la recherche ou à la compréhension des documents, ou bien à la gestion des profils des utilisateurs ; dans la seconde catégorie se trouvent l'ensemble des applications liées au marketing, qu'elles concernent l'analyse (par exemple, analyse de réponses à des enquêtes) ou la cartographie de réponses et le profiling.

L'environnement

Pour l'essentiel, ce secteur est concerné par les différents types de veille (technologique, concurrentielle, légale...). Les capacités d'analyse des technologies de la langue sont essentielles dans ce type d'applications, rendues plus difficiles et onéreuses par la multiplication des données disponibles, notamment sur Internet.

Le TAL fournissant des services "embarqués"

Le traitement automatique des langues, qu'il concerne l'oral ou l'écrit, améliore la qualité de divers services d'information, parfois sans que l'utilisateur ou l'usager en ait conscience. Ces services peuvent être "embarqués" dans une variété assez large de systèmes.

Les produits bureautiques

Les produits bureautiques, extension artificielle de nos capacités de communication, embarquent des fonctions à vocation anthropomorphe, tentant ainsi d'optimiser nos propres actions. Elles simulent alors des activités associées au langage afin de nous soulager de leurs fardeaux. On peut citer dans ce domaine les logiciels de reconnaissance optique de caractères (OCR) qui nous libèrent de la lecture et la saisie des documents, les systèmes de correction orthographique et grammaticale, qui nous soulagent des relectures fastidieuses ou de la consultation des usuels, ainsi que les outils de résumé automatique, de traduction ou les systèmes de dictée vocale...

La téléphonie, fixe ou portable

La téléphonie bénéficie grandement des traitements automatiques de la parole, qui vont au-delà de la simple possibilité de scander le nom sous lequel a été stocké le correspondant. Les systèmes de reconnaissance vocale indépendante du locuteur (IVR), notamment, alliés à des systèmes de gestion de dialogues plus ou moins sophistiqués, permettent de diminuer les coûts des centres d'appel.

Les véhicules

Les systèmes d'analyse et de synthèse vocales ont été également mis à contribution pour assister le pilote dans sa tâche, afin d'améliorer son confort ou sa sécurité. Ainsi, les premières tentatives pour embarquer un système de commande vocale dans un avion de chasse datent de 1983. Depuis, les systèmes de GPS (Global Positioning System) sont devenus des accessoires standards des véhicules et fournissent en temps réel un guidage « par la voix » de manière extrêmement fiable.

Chiffres clés du marché industriel des technologies de la langue

Selon une étude réalisée dans le cadre du projet Technolangue[1], les années 1990 ont marqué l'avènement de la société de l'information où le traitement du texte est passé dans une phase opérationnelle, et où le traitement de l'image et de la voix apparaît : l'arrivée de nouveaux supports de communication, tels que les téléphones portables, les ordinateurs portables, les bornes interactives, les systèmes embarqués, participe au développement industriel des applications. Les connexions haut débit (connexion réseau câble et satellite) facilitent l'accès à l'information. L'offre devient de plus en plus concurrentielle avec une multitude de nouveaux acteurs : plusieurs jeunes pousses en provenance de centres de recherche publics émergent sur les 9 segments d'application définis (voir schéma ci-dessous). Les pôles de recherche et les programmes européens sont des soutiens à la R&D et à l'industrialisation du secteur.

Les années 2000 marquent l'avènement de la société de la communication avec un traitement du texte, de l'image, de la voix, de la vidéo en phase industrielle. La convergence des technologies est fortement marquée dans le domaine de la communication. L'informatique, les télécommunications et l'audiovisuel sont fédérés par la numérisation. Le contenu numérique concerne désormais la voix, le son et les images : l'ère du multimédia se concrétise avec le déploiement de kiosques multimédias, des téléphones mobiles de nouvelle génération, des consoles et des terminaux de divertissement, de la domotique. Les connexions aux réseaux à haut à débit se généralisent (ADSL, fibre optique, WIFI…).

Les applications concernent les intranets documentaires multimédia, la gestion de contenu multilingue, le e-business, le vocal/multimodal, le e-learning, la traduction automatique… En plus des sociétés du CAC 40, les clients types sont les firmes multinationales, les PME intervenant dans les nouvelles technologies, le secteur public. La dynamique du marché s'oriente désormais vers une logique de la demande : intégration des outils au sein d'architectures existantes, capitalisation de l'investissement, demande permanente de l'évolution des technologies avec la généralisation du multimédia et de l'Internet mobile, et le besoin de communication multilingue.

Depuis ces dix dernières années, les fusions et acquisitions sont de plus en plus fréquentes et s'expliquent par :

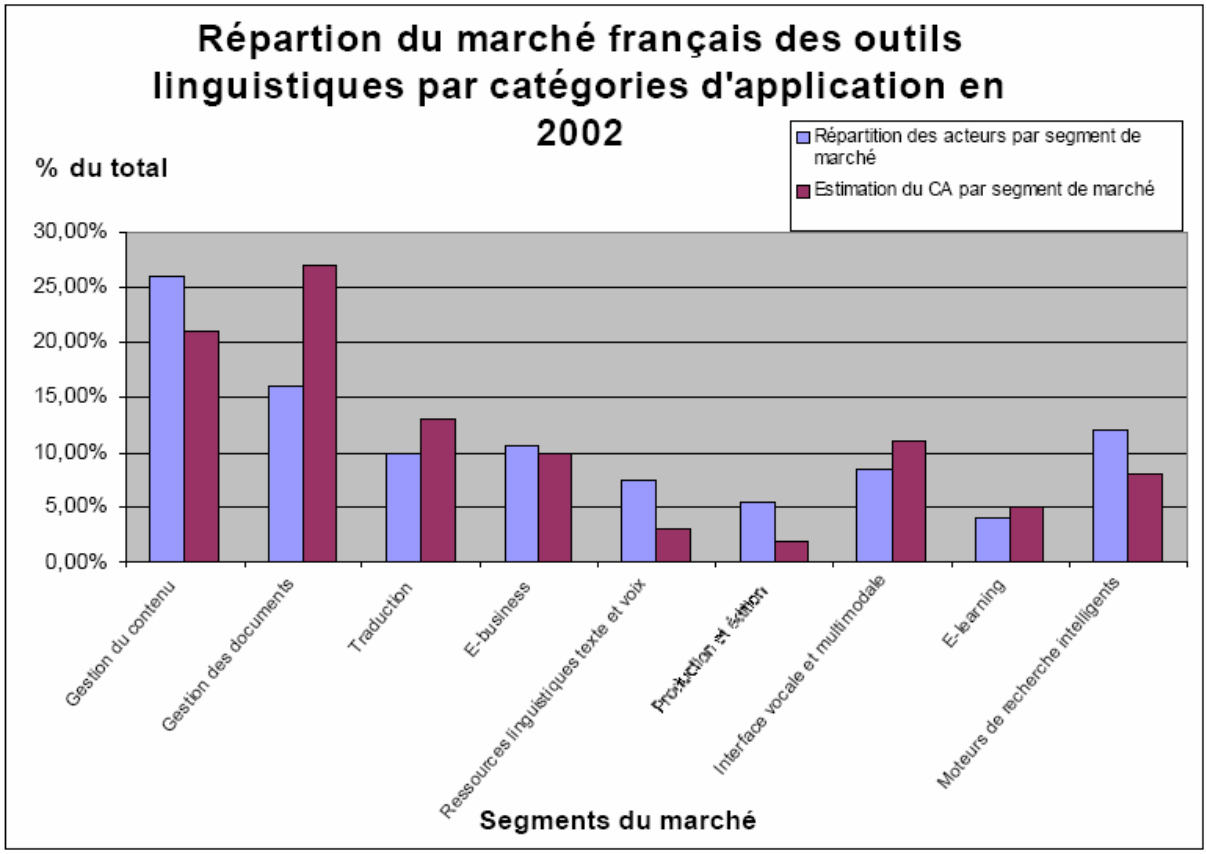
- la frénésie du marché de l'Internet dans les années 1990 avec la course au premier entrant qui a permis de développer des partenariats et de valider des fusions entre acteurs concurrents ;
- le marasme économique du secteur des NTIC au début des années 2000 avec l'éclatement de la bulle Internet et la perte de valeur de plusieurs acteurs de l'offre ;
- le positionnement stratégique multinational des offreurs de technologie sur le marché des outils linguistiques. Ce marché devient global (interconnexion des acteurs européens et américains), comme en témoigne la pénétration des acteurs américains en Europe ;
- la volonté de contrôler des technologies stratégiques pour la sécurité ou la compétitivité nationales.

La demande d'accès à l'information s'oriente vers des solutions en langage naturel qui autorisent l'interrogation multilingue. La diffusion de l'information avec les technologies associées a engendré l'éclatement de la fonction « Gestion de l'Information » : les systèmes de gestion sont orientés vers les utilisateurs finals et non plus ciblés vers les

[1] Cette section reprend largement l'étude *Technologies de la langue en Europe : marché et tendances* réalisée par le Bureau Van Dijk, à la demande du Ministère de la recherche dans le cadre du programme Technolangue et disponible en ligne sur le site www.technolangue.net.

seuls experts de la documentation. L éparpillement de la demande pose le problème de l identification des besoins et des attentes des différents groupes d utilisateurs.

En 2002, 377 offreurs de technologies ou services linguistiques étaient présents sur au moins un des 9 segments d applications retenus dans le cadre de cette étude.



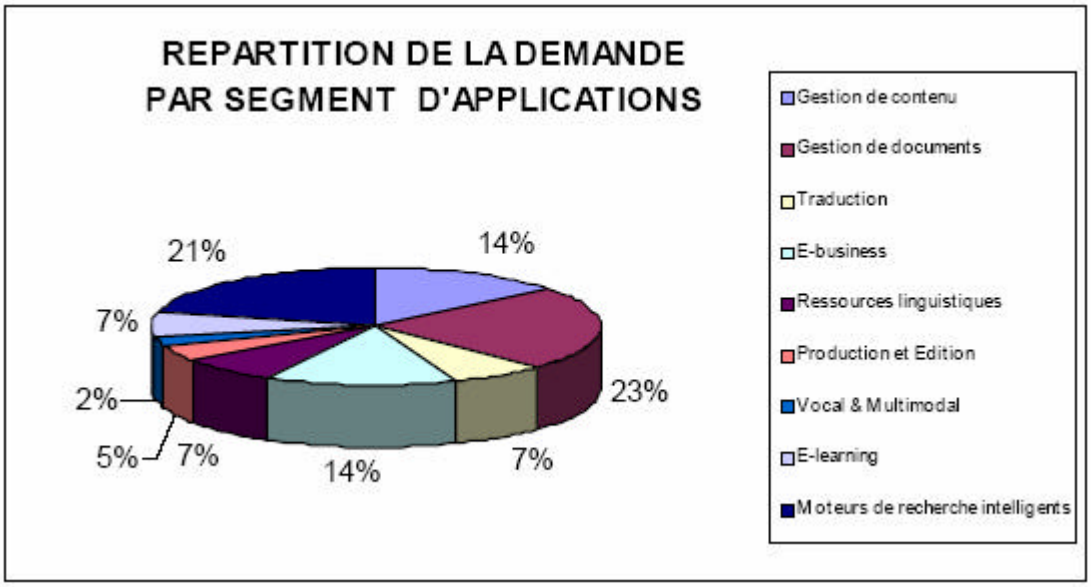
L offre se répartit entre le traitement du texte (80%) et le traitement de la voix (20%). Le marché européen était évalué, en 2002, à 510 millions d euros, dont 60% pour le Royaume-Uni, la France, l Allemagne et l Italie.

96 sociétés françaises ont été identifiées dans le périmètre de l ingénierie linguistique, dégageant une chiffre d affaire de 77 Millions d euros en 2002, soit 15%du marché européen, la France se plaçant ainsi au deuxième rang des pays de la zone.

Du coté de la demande, les secteurs les plus représentatifs sont l industrie, y compris Pharmacie et santé (36%), les services/banques assurances finances (20%), le public (15%) et les transports/tourisme (11%).

Les sociétés du CAC 40 et la plupart des administrations publiques représentent la majorité des entités demandeuses de solutions de traitement du langage.

La répartition de la demande exprimée par segments d'applications se présente ainsi :

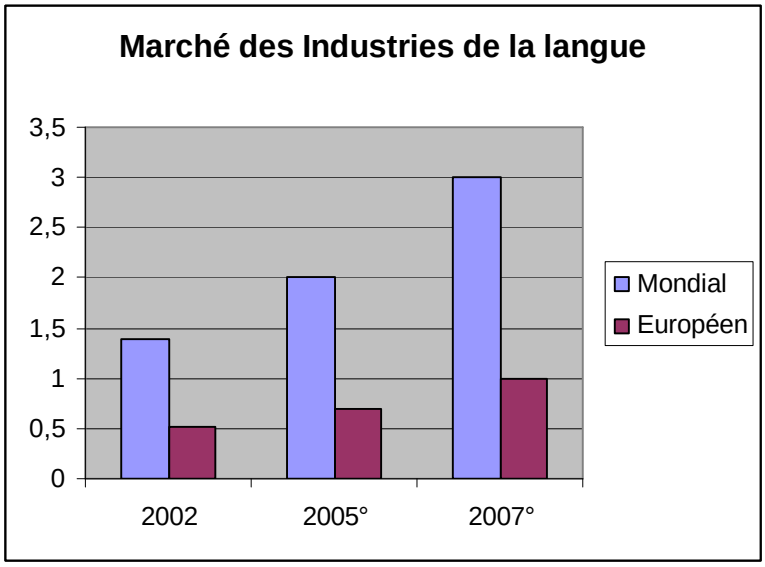


La répartition de la demande associée aux solutions mises en place est représentée par des applications dédiées à des groupes d'utilisateurs spécifiques au sein des sociétés : 52% des utilisations concernent une cible utilisateur de 100 à 1000 personnes, 12 % une utilisation pour plus de 1000 personnes. Les applications sont de plus en plus orientées réseau (Internet, Intranet, Extranet) avec des architectures client/serveur. Elles ciblent un groupe d'utilisateurs de plus en plus large, même si encore 18% des applications mises en place concerne un nombre restreint d'utilisateurs (nombre d'utilisateurs inférieurs à 100) lié à l'étroitesse et à la spécialisation de la cible.

Les applications mises en place en univers industriel (hors usage des particuliers) impliquent un processus décisionnel au sein de l'entreprise variant de 12 à 18 mois en moyenne. Plus de 60% des applications déployées datent de moins de 5 ans. Le marché de la demande est encore jeune dans la mesure où les outils déjà mis en place ne permettent pas de capitaliser intégralement la maîtrise de l'information dans l'entreprise. En ce qui concerne les évolutions et les attentes de la demande, elles sont liées pour 82% à l'accroissement du volume des données à traiter et de la montée en puissance des réseaux. Ce qui implique que près de 35% de la demande est intéressé par les systèmes de résumés automatiques, et 41% par la veille automatisée sur Internet.

En guise de conclusion, on peut estimer que ce marché mondial des technologies de la langue devrait dépasser les 2 milliards d'euros en 2005 et dépasser les 3 milliards d'euros en 2007. Le marché européen devrait atteindre 1 milliard d'euros en 2007. La progression de l'usage des NTIC laisse présager un élargissement du marché des outils linguistiques vers le grand public. Le besoin est ressenti d'engager des actions marketing pour optimiser l'adéquation de l'offre à la demande.

Ainsi, selon cette étude, il ressort que le marché des industries de la langue devrait continuer son évolution spectaculaire.



°prévisions

Une brève introduction au Traitement automatique des textes

Alain Couillault

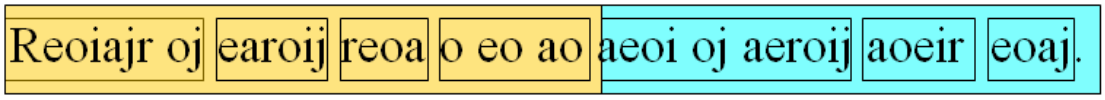
Ce chapitre se veut une brève introduction au traitement automatique de la langue et des différents niveaux d'analyse d'information textuelle, destiné au béotien et, à ce titre, volontairement simplificateur.

Pour illustrer notre propos, imaginons que nous soit donnée la tâche d'analyser un bref texte dans une langue qui nous serait inconnue, représenté, par exemple, par la suite de caractères suivante :

Reoiajr oj earoij reoa o eo ao aeoi oj aeroij aoeir eoaj.

Découper

Pour démêler ce texte sibyllin, il nous faut tout d'abord en connaître les segments qui la composent, c'est à dire y reconnaître ce que l'on appelle habituellement des phrases ou des mots. C'est le rôle des outils de **segmentation**, dont la complexité peut varier depuis la simple reconnaissance de caractères d'espace ou de ponctuations, à l'utilisation de dictionnaires complets. Un bref inventaire des cas à traiter met assez vite en évidence les limitations du premier. Un segmenteur pourra ainsi proposer le découpage en phrases et mots suivants pour le texte ci-dessus :



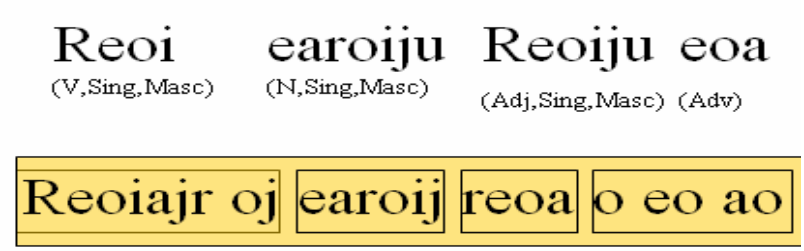
Le segmenteur nous indique ainsi que le texte est composé de deux phrases, la première étant constituée de quatre mots et la seconde de trois. Le point n'a ici pas été étiqueté comme un mot.

Notons que le découpage d'une langue en segments, dans le cadre du TAL, dépend des objectifs de cette segmentation et des ressources dont on souhaite disposer pour effectuer l'analyse.

Ce niveau de traitement est souvent considéré comme suffisant pour différentes applications du TAL, notamment celles qui concernent les cas simples de recherche plein texte, la fouille de texte ou la cartographie d'information.

Etiqueter

Evidemment, la langue n'étant pas une suite de borborygmes, l'ensemble des mots qui constituent une phrase ne sont pas équivalents. Ce sont en général des formes particulières d'un certain vocabulaire qui sont porteurs d'informations telles que le nombre, le genre, la personne... **L'étiquetage** consiste à reconnaître ces informations. Un étiqueteur pourra, par exemple, proposer l'analyse suivante pour la première phrase de l'exemple ci-dessus :

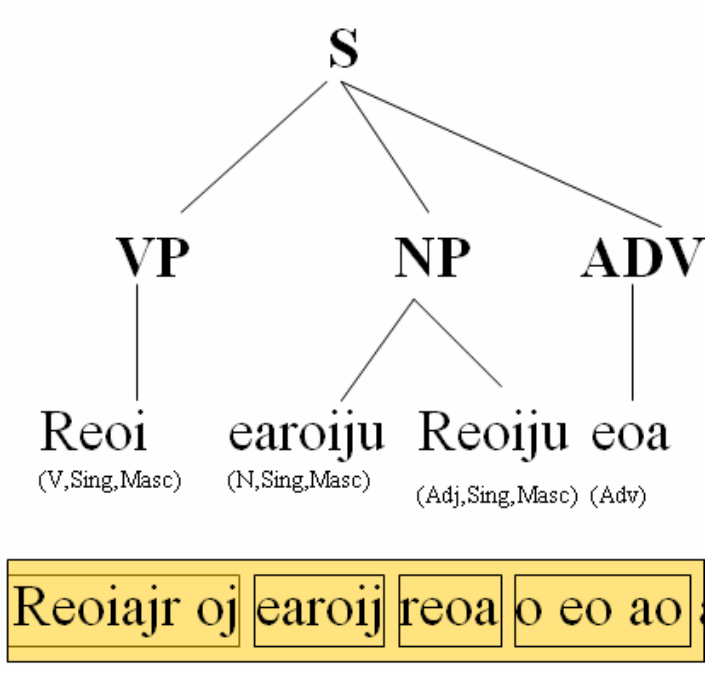


Ainsi, dans cet exemple, la chaîne ‘Reoiajr oj ’’ est reconnue comme une forme du verbe ‘reoi ’’, au singulier et au masculin (dans notre Volapuk, les verbes portent la marque du genre). Notons que l étiqueteur fictif utilisé ici effectue à la fois la reconnaissance de ce qu on peut appeler les mots du dictionnaire et l attribution des étiquettes. Cette étape peut d ailleurs être combinée avec l étape précédente, la segmentation en mots étant accomplie par le même module. C est également à cette étape que peut s effectuer la lemmatisation, qui consiste à reconnaître la forme canonique d un mot.

Là encore, la manière de procéder à l étiquetage, le choix des étiquettes et les informations fournies par le module dépendent largement de choix scientifiques, méthodologiques et applicatifs. L étiquetage est assez souvent une étape vers des traitements plus complexes. Il peut être suffisant pour des environnements de recherche plein texte avancés.

Reconnaître la structure

Les mots entretiennent entre eux des relations de différentes natures, comme celles qui existent en un verbe, son sujet et ses compléments, un nom avec son adjectif ou son déterminant. L **analyse syntaxique** a pour rôle de reconnaître ces relations, représentées ci-dessous sous la forme d un arbre.

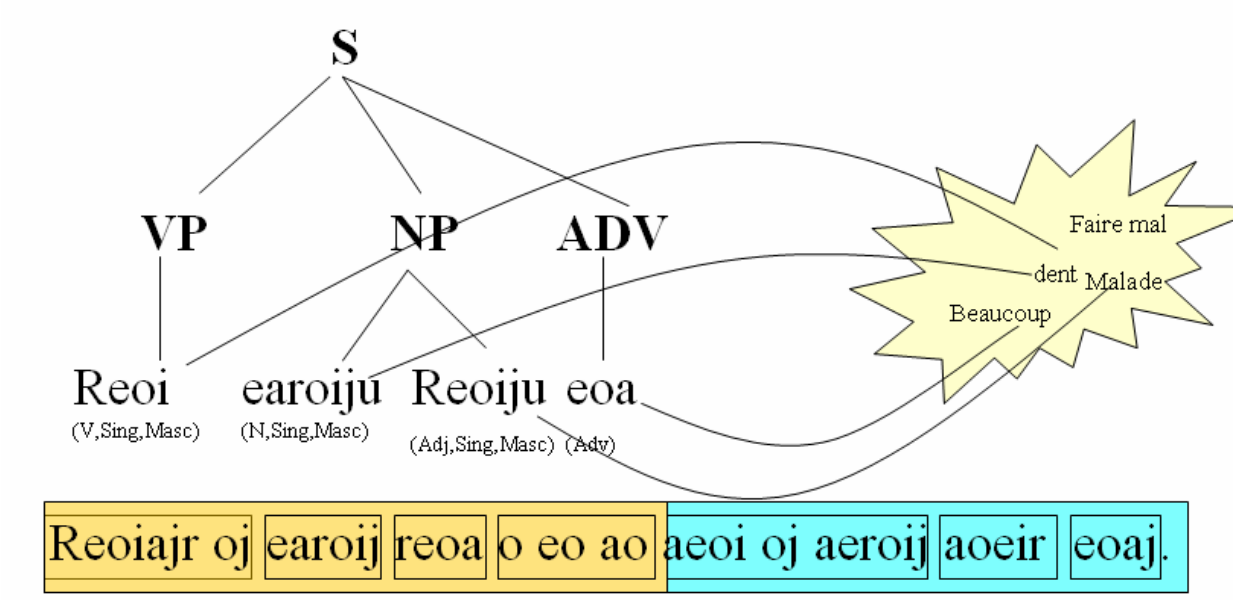


Cet arbre nous indique que la suite « earoij reoa » est un groupe nominal[2] constitué d'un nom et de son adjectif. Ce groupe nominal est lui-même le sujet du verbe, le sens de la phrase étant modifié par l'adverbe « e eo ao ». Evidemment, nous ne savons encore rien du sens de la phrase. Il n'est pas toujours nécessaire, pour des besoins applicatifs, d'analyser l'ensemble d'une phrase, voire même de décider entre une analyse ou une autre lorsque la phrase paraît ambiguë à un système d'analyse. Reconnaître des expressions, et notamment des groupes nominaux, est utile à des applications de diverses natures : reconnaissance ou élaboration de terminologies, recherche plein texte, reconnaissance d'entités nommées (i.e. reconnaître des noms de lieu, de personnes, des dates…), la fouille ou le résumé de textes…

Evaluer le sens

La notion de *compréhension* a fait couler beaucoup d'encre et d'octets, et nous n'y reviendrons pas ici. Disons simplement qu'on peut considérer que la mise en relation d'un texte avec une structure représentant le sens des mots (ce qu'on appelle généralement une *base de connaissances*) constitue une analyse sémantique. Ces bases de connaissances organisent les mots entre eux, généralement en les associant à des concepts[3] et en décrivant la nature des relations qui les unit. Le terme désormais consacré d'ontologie[4], tel qu'il est utilisé dans le cadre du W3C, décrit lui aussi une telle structure.

Si une base de connaissances existe pour notre fameuse langue de travail, une analyse sémantique de la phrase pourrait être représentée par la figure ci-dessous :



Ce schéma nous dit que le verbe de la phrase a un rapport avec la douleur, que le nom dénote cet *'organe dur, blanchâtre, généralement composé d'une couronne libre et d'une (ou de) racine(s) implantée(s) dans la cavité buccale et, plus particulièrement, sur le*

[2] « groupe nominal » fait partie de ces expressions mieux connues des jeunes générations que des anciennes, à coté de « console de jeux », ou « mp3 » mais pour des raisons différentes.

[3] Notons que dans la littérature, ce terme est malheureusement ambigu. A la suite de Salton, les adeptes des approches statistiques en recherche d'information l'utilisent pour dénoter ce que l'on a ici appelé « expression », un autre courant, plus traditionnel, l'utilise au sens classique défini par Platon.

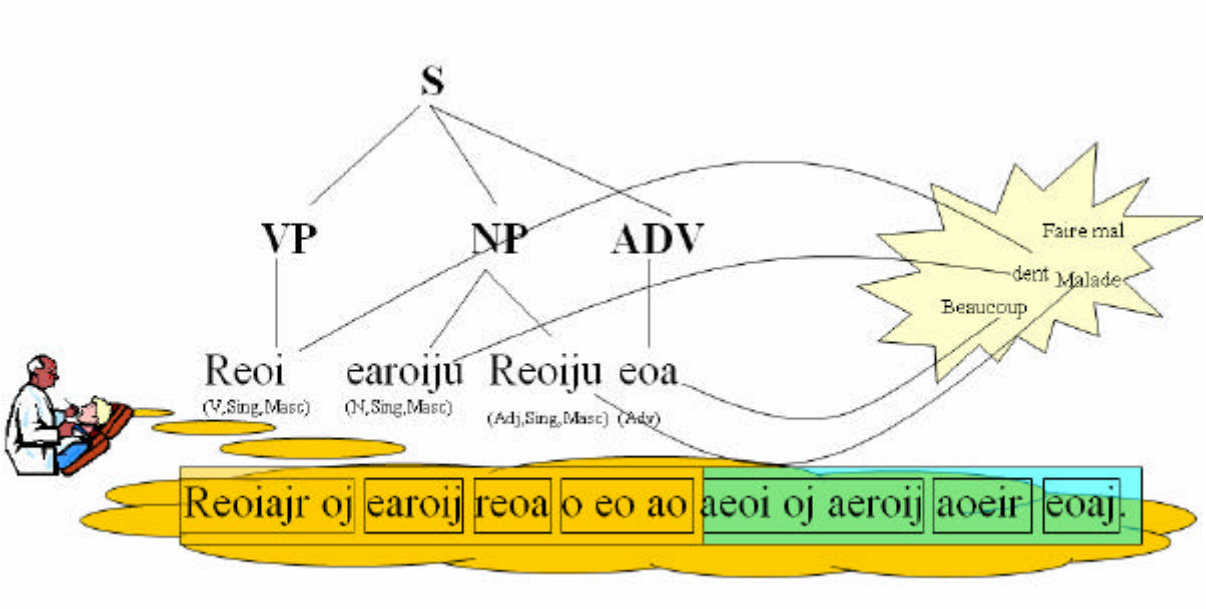
[4] Ce terme, de façon assez similaire, est également ambigu.

rebord libre des maxillaires, et destiné à saisir, retenir et broyer les aliments [5]''. Les relations entre les termes ne sont pas représentées dans cette pseudo base de connaissances, mais un vrai système pourrait inclure des liens avec des synonymes (par exemple, en français « être douloureux » pour le verbe), des termes génériques ou spécifiques (par exemple, « molaire » pour « dent ») ou d'autres types de relations. Elle peut également décrire des relations avec des termes d'autres langues.

Ce niveau d'analyse est utile dans deux grands champs applicatifs. Celui de la recherche d'information où, pour l'essentiel, il continue de manière automatique l'usage traditionnel des thésaurus ou des listes d'autorité. Celui également de la traduction automatique, où il permet d'établir le lien entre plusieurs langues.

Mettre en contexte

Nous percevons, à ce stade le sens général de la phrase[6], il faut désormais la mettre en contexte pour évaluer son utilité. Ainsi, la phrase peut être prononcée par un patient chez son dentiste (comme illustré ci-dessous), il s'agit alors d'un appel pressant à faire le nécessaire pour que le dentiste fasse son office. Prononcée dans d'autres circonstances, par exemple dans une pharmacie ou auprès des renseignements téléphoniques. Ce domaine de la mise en contexte des phrases est, pour les linguistes, celui de la pragmatique, qui commence à poindre dans le traitement automatique des langues, par le biais d'une plus grande prise en compte des particularités de l'utilisateur (un terme n'a pas le même sens pour un utilisateur ou un autre) ou la recherche d'experts.



Et dans l'autre sens

Ce léger survol s'est intéressé à l'analyse des suites de caractères, les mêmes niveaux peuvent être décrits pour ce que l'on appelle la génération, que ce soit pour des besoins de génération de résumés, de traductions ou de génération de textes, par exemple à partir de bases de données.

[5] Source : Le Trésor de la Langue Française

[6] Soyons modestes, nous le percevons parce que doués de facultés humaines, ce n'est pas nécessairement le cas de la « machine ».

Quelques applications en détail

L'analyse des langues constitue un élément important, voire essentiel, pour certaines applications industrielles. Les exemples ci-dessous sont tirés de projets réels. Pour chacun sont décrits le contexte de l'entreprise, ses besoins fonctionnels, les fonctionnalités et les retours de la solution mise en oeuvre.

Veille

Alain Garnier

Mots-clés : Veille, Intelligence économique & stratégique, Stratégie.

Contexte, Cas Entreprise

Les grandes entreprises et les grosses PME sont aujourd'hui confrontées à une compétition accrue qui les amène à industrialiser leur processus de surveillance du monde extérieur, en particulier de leurs concurrents. Ce processus s'inscrit dans un ensemble de problématiques très large qui englobent la technologie, le positionnement produit ou des questions géostratégiques. Ces responsabilités sont généralement regroupées au sein de cellules dédiées à cette activité, dites "de veille" ou "d'intelligence économique".

Description du besoin métier

Une cellule de veille a trois missions principales :

La première consiste à suivre sur le long terme les axes de veille qui ont été déterminés par la stratégie de l'entreprise. Elle doit donc constituer un corpus représentatif de l'information disponible autour de ces thèmes et en assurer un suivi puis une synthèse sous forme d'analyses.
La deuxième consiste à être en mesure de répondre à tout moment à une crise afin de fournir aux décideurs les bonnes informations pour prendre une décision.
La troisième consiste à irriguer l'organisation de ces pratiques afin que l'intelligence économique devienne une habitude de travail pour tous.

Description du processus actuel / chaîne de traitement

Aujourd'hui, le travail consiste pour la majeure partie à collecter l'information disponible sous de multiples formats (papier, CD-ROM, web, bases de données…) afin de constituer des dossiers de veille. Viennent ensuite les missions d'analyse qui sont réalisées « dans le temps qui reste ».

Description des problèmes actuels

Ce processus manuel est très coûteux en temps et en argent.
Les conséquences sont notamment :

- L'information disponible trop volumineuse conduit à une vision partielle de la réalité,
- La réactivité pour traiter un sujet est «lente »,
- L'analyse est réduite à une portion congrue au détriment du travail fastidieux de collecte.

Analyse

- Les technologies TAL permettent d'automatiser la gestion de ce flux d'information «au delà de la taille humaine »
- Le TAL permet également de garantir un niveau de fiabilité dans le filtrage qui répond aux exigences de la problématique.

C'est donc une technologie qui change, au niveau qualitatif, la façon dont le problème est abordé.

Identification du point d'intervention TAL

Le TAL intervient à tous les niveaux de la chaîne de mise en valeur de l'information.

Tout d'abord, en phase de collecte, la compréhension en profondeur des documents permet un filtrage efficace pour passer, sur un sujet donné, des quelques milliards de pages sur le web par exemple aux 10 000 documents utiles.

Ensuite, le TAL permet de ranger l'information dans des catégories très fines auxquelles les utilisateurs vont s'abonner afin de ne recevoir que l'information utile qui les concerne.

Enfin, le TAL, par sa capacité d'extraction d'information riche (nom de personne, de sociétés etc...), va permettre une analyse fine autour de l'information tout en donnant un accès direct à l'information utile dans un document.

Description de la technologie TAL applicable

Les trois phases du système de veille s'appuient sur une technologie TAL de type analyse sémantique.

Description du gain qualitatif / quantitatif attendu

- D'un point de vue qualitatif :
  - le système automatique permet de faire travailler une équipe entière et géographiquement dispersée sur un même dossier ;
  - le système permet d'obtenir en temps réel des informations sélectionnées par les profils de veille ;
  - le système permet de constituer une mémoire collective autour des sujets stratégiques pour l'entreprise ;
  - le système ne « passe pas à côté » de modifications des sites ou des sources d'information.
- D'un point de vue quantitatif :
  - Le système gère un volume cent à mille fois supérieur à la même chose effectuée manuellement
  - Le temps passé à l'analyse est multiplié par deux

Déploiement et mise en œuvre

Évolution de la chaîne de traitement

Installation

Le logiciel s'installe en une journée pour mettre en place la structure du serveur. Un travail de paramétrage DSI est indispensable pour donner au système et aux utilisateurs les droits d'accès requis.

Utilisation

1. Définition des profils de veille : Il s'agit de paramétrer des sources d'information pertinentes pour les axes de veilles choisis. Le système TAL permet de filtrer au sein des volumes d'informations ceux qui concernent directement les axes choisis.
2. Classement de l'information : chaque source peut être classée automatiquement par des technologies sémantiques pour enrichir des catégories de veille très fines.
3. Publication : le résultat de cette mise en valeur de l'information est soit publié dans un portail, soit envoyé en mode « push » vers les utilisateurs

Description du traitement TAL appliqué

Le TAL consiste principalement à fournir une « vision » de l'information qui va au-delà du texte. Par exemple, un concurrent est défini dans le système de manière sémantique, ce qui permet par la suite de filtrer selon l'axe « concurrentiel ». Le TAL permet donc de regrouper et d'affiner les recherches, filtrages et classements.

Évaluation ROI

Coûts et délais de mise en œuvre

Le coût de mise en œuvre se décompose en deux parties :

1. l'investissement initial consiste en du matériel, du logiciel et du service. Pour le matériel, un serveur sous Windows est préconisé. Le logiciel a un coût d'entrée de gamme de l'ordre de 25 à 50k€. Le service nécessite une vingtaine de jours au départ (formation, mise en place). Le service est en général opérationnel en quelques mois.
2. le coût de maintien du service est essentiellement un coût humain constitué d'une part par la DSI associée au système serveur et d'autre part par les personnes qui administrent la solution.

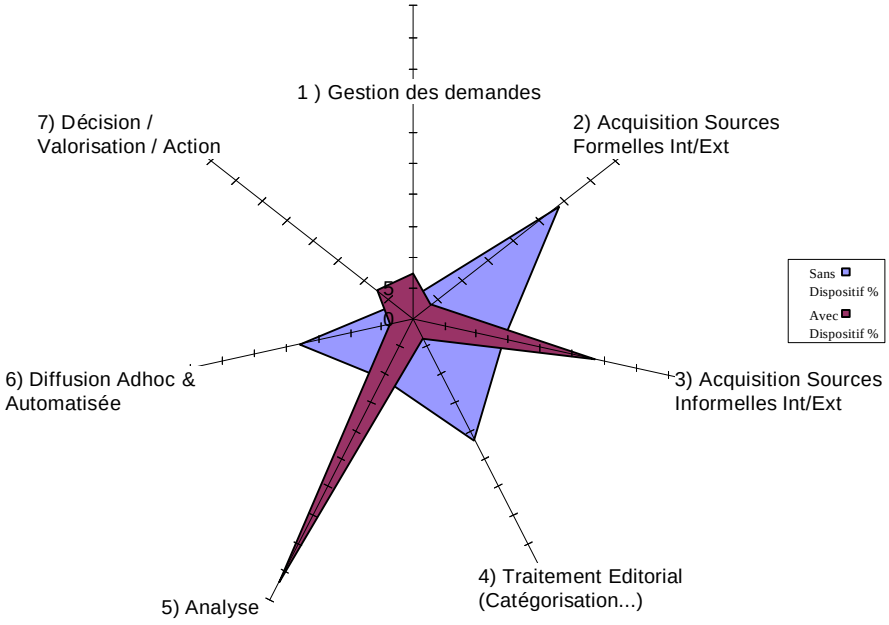
Gains fonctionnels / opérationnels

Nouvelles fonctionnalités fournies par l'outil :

- Automatisation de la collecte et de la surveillance notamment du web
- Classement automatique de grands volumes d'informations
- Moteur de recherche sémantique multilingue
- Push automatique ciblé à coût très réduit
- Outil de lecture rapide de corpus

Gains opérationnels :

- Gain pour les veilleurs de 50% de temps de collecte et traitement repositionné pour l'analyse et la diffusion
- Gain de productivité de l'ordre de 100% (doublement du nombre d'axes de veille à effectif constant).
- Gain pour les utilisateurs finaux qui peuvent directement accéder à l'information utile sans perdre de «temps » sur Google.



Retours Utilisateurs

Le client utilise la solution depuis maintenant 3 ans et en est pleinement satisfait. Le nombre de profils de veille a doublé et le nombre d'utilisateurs a été multiplié par dix.

La cellule traite désormais un nombre de demandes beaucoup plus important que par le passé, sans changements organisationnels.

Enfin, la diffusion au sein du portail de l'entreprise a permis une meilleure visibilité de l'activité.

Veille en Intelligence Economique

Guillaume Mazières
Sylvie Guillemin-Lanne

Mots-clés : Veille, Intelligence économique & stratégique, Stratégie.

Contexte, Cas Entreprise

La Direction Veille Information Archives, au sein de la Direction des Ressources Humaines et de la Communication, est en charge de l'information pour l'ensemble du groupe spécialisé dans le domaine pétrolifère, à travers ses différents secteurs d'activité.

Description du besoin métier

Cette cellule de veille scrute et analyse l'information géopolitique, technique, financière, sociale, micro et macroéconomique de son environnement. Elle a pour mission :

- de fournir une information de qualité
- de mettre en place des processus opérationnels de collecte et d'analyse.

Description du processus actuel / chaîne de traitement

Afin de répondre aux fréquentes demandes de dossiers thématiques relatifs à leur activité, la Direction Veille prend en charge les actions de collecte et d'analyse. Les sources d'information sont nombreuses et variées. Elles concernent à la fois les fournisseurs de données de presse, tels que Factiva ou Lexis Nexis qui fournissent de l'information segmentée selon différentes thématiques, et les périodiques disponibles sur le portail du groupe, les communiqués de presse ou encore les rapports d'analyse sectorielle de banque.

La Direction Veille a donc un besoin urgent d'automatisation et de rapidité d'accès à l'information afin d'être capable de restituer une information triée et organisée en vue d'une exploitation efficace.

Description des problèmes actuels

Aujourd'hui, l'accroissement des volumes d'information à traiter, dû à la quantité des sources disponibles et à la diversité des demandes internes, rend les traitements manuels de lecture et de synthèse difficiles et très coûteux.

Analyse

Identification du point d'intervention TAL

La solution TAL intervient sitôt après la collecte des documents. Le TAL permet de procéder à une analyse textuelle de tous les documents et d'extraire de ceux-ci l'information de veille jugée pertinente par le client. Seront ainsi extraits les noms de sociétés qui intéressent le client et toutes les informations afférentes aux actions de ces sociétés.

Description de la technologie TAL applicable

La solution d'extraction d'information enchaîne trois étapes d'analyse linguistique :

- analyse morpho-syntaxique : affectation à chaque mot d'un document d'une catégorie grammaticale (nom, adjectif, verbe…) assortie de traits morpho-syntaxiques (genre, nombre),
- lemmatisation : retour à la forme canonique de chaque mot (singulier pour un pluriel, infinitif pour un verbe conjugué) pour qu'il soit reconnu indépendamment de sa forme fléchie,
- extraction de connaissance (exécution des règles d'extraction) : identification des entités (noms de personnes, noms de compagnies, valeurs, dates, lieux, etc.), reconnaissance des relations entre les entités (relation d'achat, de cause à effet entre 2 sociétés, etc.).

Description du gain qualitatif / quantitatif attendu

- D'un point de vue qualitatif :
  - Permettre une analyse fine et homogène des documents
  - Centraliser la connaissance pour éviter que les documents soient analysés plusieurs fois
- D'un point de vue quantitatif :
  - Réduire de 50% le temps consacré au quotidien par les veilleurs à l'analyse de documents stratégiques

Déploiement et mise en œuvre

Évolution de la chaîne de traitement

Pour atteindre ces objectifs, il a été développé une application TAL qui repose sur le couplage d'un serveur d'extraction d'information à un composant sémantique dédié à l'Intelligence Economique. Un autre composant adapté au monde de l'Industrie pétrolière regroupe l'ensemble des lexiques et des règles d'extraction liée à cette thématique :

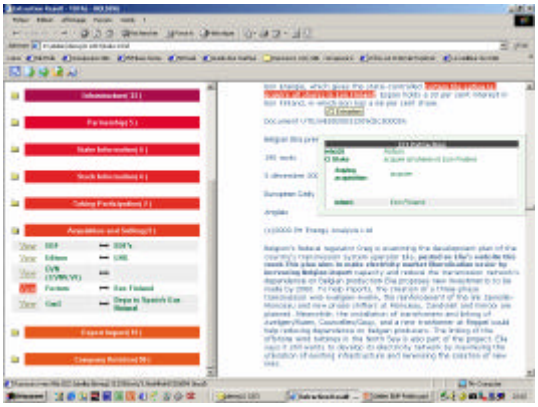
- Le serveur d'extraction d'information procède à l'analyse des mots d'un document et renvoie pour chacun d'eux leur lemme et leur catégorie grammaticale

- A l'aide de ces informations, combinées à celles contenues dans les composants sémantiques, il procède à l'extraction d'information.

Utilisation

L'utilisation de cette solution d'Intelligence Economique permet d'extraire instantanément des flux de presse des données concernant des données financières (chiffre d'affaires, rentabilité, croissance), commerciales (parts de marché, nombre de clients), boursières (capitalisation, tendances), mais également toutes les informations concernant les prises de participation, les fusions, les acquisitions, les joint-ventures, les axes de recherche, les innovations…

La Direction Veille met ensuite cette information, actualisée quotidiennement, à disposition de ses clients internes sur son portail groupe.

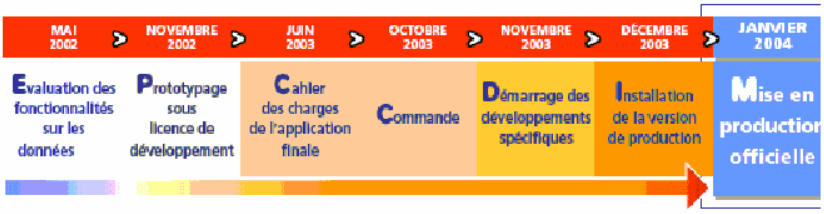


Description du traitement TAL appliqué

Analyse morpho-syntaxique, extraction d Information, composant sémantique

Évaluation ROI

Coûts et délais de mise en œuvre



Gains fonctionnels / opérationnels

Nouvelles fonctionnalités fournies par l'outil :

- Gain en fiabilité : Le serveur d'extraction d'information identifie les différents concepts avec précision, le taux de pertinence des informations extraites dépassant les 90%.
- Gain en flexibilité : Les utilisateurs ont la possibilité de travailler par thématique ou par suivi de société.

Gains opérationnels :

- Gain de temps : La lecture par concept est très rapide et il ne faut plus que 3 clics pour accéder à une information pertinente.
- Pertinence et flexibilité de la solution proposée : la Skill Cartridge™ Competitive Intelligence correspondait bien aux besoins du client, ne nécessitant que l'ajout de quelques concepts spécifiques (infrastructure pétrolière, chiffres d'import / export, co-marketing…).
- Capacité d'évolution de la solution permettant au client d'envisager rapidement l'ajout de nouveaux axes d'analyse stratégique et l'intégration de nouvelles langues.
- Gain de productivité : Selon le métier, chaque personne passe entre 15 minutes et une heure à traiter de l'information stratégique. La solution d'extraction automatisée permet au minimum un gain de temps de 50%. Pour 5000 utilisateurs dans le groupe, les gains de productivité représentent plus de 100 jours hommes cumulés par jour d'utilisation du logiciel.

Retours Utilisateurs

Pertinence de l'accès à l'information textuelle proposé pour l'intelligence économique.

Portail

Hugues de Mazancourt

Mot-clé : Fourniture de contenu en ligne.

Contexte, Cas Entreprise

Société de services d'aide à l'exportation française (plus de 5 000 entreprises clientes). La société gère un site Web portail présentant un contenu « interne » structuré comme suit :

1. des fiches « pays » décrivant l'activité économique, les principales formalités, les écueils à l'exportation pour la quasi totalité des pays du monde,
2. des fiches décrivant les foires et salons, ainsi que les organisations professionnelles dans le monde,
3. des fiches « opportunités et tendances », rédigées par les analystes en interne, qui décrivent les tendances économiques pour un produit ou secteur d'activité dans un pays (ou groupe de pays),
4. des fiches « études de marché », qui sont collectées par les analystes depuis des sites publics (gouvernementaux ou autres).

De plus, le portail présente un contenu « externe » qui est le résultat d'une surveillance de 2 000 sites ciblés par la société et utiles dans le contexte d'une étude sur l'export (sites de douanes, de marketing, d'économie en général pour un pays ou un secteur donné).

Le renouvellement du contenu interne est d'environ 50 fiches « opportunités et tendances » et 50 études de marché par semaine. Le contenu externe évolue au rythme de ses éditeurs et recense environ 600 000 pages.

Le site portail compte environ 25 000 visiteurs par mois. Il diffuse des alertes à ses abonnés (service payant), avec le même type de recherche que pour la consultation du site.

Description du besoin métier

Le besoin est de fournir un portail simple d'accès aux divers contenus. Si un certain nombre d'éléments (fiches pays, par exemple), peuvent être présentés avec des techniques standard de portail, une fonction de recherche intelligente est nécessaire pour les autres (cf. ci-dessous). Le site est à la fois fournisseur de revenus en tant que tel (certains contenus sont payants) et un point d'entrée vers des services spécialisés proposés par la société.

Description du processus actuel / chaîne de traitement

Le contenu interne est remis à jour une fois par semaine par les analystes de la société. Le contenu externe est mis à jour par un « crawling » régulier des 2 000 sites. L'ensemble des données est stocké dans une base relationnelle. Une interface simple permet d'entrer ces données. Elles sont périodiquement transférées sur le serveur, dans une base de données.

Description des problèmes actuels

Une fonction de recherche avancée est nécessaire pour mettre en valeur les contenus à la fois externe et interne. En effet, le contenu, même riche, est d'un volume relativement « petit » comparé aux grandes bases de données commerciales ou au Web. On se heurte donc aux problèmes classiques de bruit et (surtout) de silence lors d'une interrogation par mots-clés.

De plus, comme le contenu est bilingue (français et anglais, mais anglais à 80%), l'utilisateur doit faire plusieurs requêtes en fonction des documents interrogés, éventuellement toutes dans des langues différentes de la sienne, dans la mesure où le site existe en version francophone, néerlandophone ou anglophone.

Analyse

On va donc opter pour un système qui enrichisse les requêtes de synonymes et mots proches à partir d'un dictionnaire général avec une orientation « business » et éventuellement augmenté de terminologie spécifique, doté de capacités multilingues (« cross-language »).

Identification du point d'intervention TAL

La solution TAL, du point de vue de l'utilisateur, n'intervient qu'au travers de la fonction « rechercher » du portail. Elle est la plus transparente possible.

Dans les faits, les contenus textuels sont indexés via le système TAL qui fournit également la solution de recherche dans ces contenus.

La fonction d'alerte utilisera la même recherche (avec un filtre sur la date des documents).

Description de la technologie TAL applicable

- enrichissement de requête
- expansion à d'autres langues,
- détection de la langue des documents (au moment de l'indexation),
- « crawling » et conversion de formats (technologies sans lesquelles les solutions TAL ne peuvent être mises en place).

Description du gain qualitatif / quantitatif attendu

Le gain qualitatif s'exprime en précision ou rappel sur les documents obtenus par la recherche. Typiquement, sur ce type d'applications, le gain va jusqu'à multiplier par 5 ces indicateurs. L'objectif de recette du projet était 80% de précision à 10 réponses sur un corpus de requêtes pré-établi (par rapport au contenu interne).

Le gain quantitatif s'évalue par la qualité des réponses et aussi par le fait que le système utilise une seule requête pour interroger des documents dans diverses langues. Une requête sur les « machines agricoles » trouvera des documents anglais parlant de « harvester ».

Une fonction de « dialogue » permet à l'utilisateur de voir comment sa requête a été enrichie et de raffiner les termes utilisés pour la recherche.

Déploiement et mise en œuvre

Évolution de la chaîne de traitement

L'interrogation est appelée sous la forme d'un Web Service, qui renvoie les éléments nécessaires (identifiants de documents ou mots utilisés pour la recherche). L'intégration dans le site (développé en ASP) est donc particulièrement souple et aisée.

La synchronisation des données vers le système d'indexation se fait par FTP, chaque nuit.

Description du traitement TAL appliqué

- analyse syntaxico-sémantique de la requête,
- désambiguïsation sémantique,
- expansion via un dictionnaire multilingue,
- interrogation de la base documentaire avec la requête enrichie,
- utilisation de techniques de recherche linguistique dans des classifications pour identifier le secteur d'activité (dans la nomenclature du client) à partir de cette même requête.

Évaluation ROI

Coûts et délais de mise en œuvre

Investissement : coût de la licence du produit + intégration de la fonction de recherche dans le portail (simple).

L'administration est très simple (synchronisation des fonds) et n'a pas entraîné de surcoût.

Gains fonctionnels / opérationnels

Les gains s'expriment en qualité de recherche.

Pour la société, le fait de disposer d'une interrogation cross-language va lui permettre d'ouvrir des sites présentant le même fonds documentaire dans d'autres pays.

Retours Utilisateurs

Pas de retour direct de la part de la société. De plus, le service s'est ouvert avec une nouvelle version du portail qui comprenait d'autres améliorations sur le fond et le contenu. Il est donc difficile de mesurer l'impact de la technologie TAL elle-même.

Toutefois, un bon indicateur est que le service est passé, après quelques mois d'activité, d'un mode « gratuit avec inscription » à un mode « payant », ce qui implique que le nombre de clients et la qualité du service étaient suffisants pour faire accepter de le faire payer.

Classification automatique

Guillaume Mazières
Sylvie Guillemin-Lanne

Mots-clés :Text Mining, Classification automatique, Plan de classement, Verbatims clients

Contexte, Cas Entreprise

Les constructeurs automobiles surveillent de près la qualité perçue par leurs clients en recueillant, par le biais d'enquêtes, leurs sentiments sur les véhicules qu'ils viennent d'acquérir. Les propriétaires de nouveaux modèles sont contactés systématiquement 3 mois après leur acquisition. Ces enquêtes sont l'occasion pour les constructeurs de collecter des informations qualitatives et stratégiques.

Description du besoin métier

Le Département Qualité, responsable des enquêtes, est chargé de fournir des résultats pertinents et organisés aux différentes entités du groupe intervenant tant au niveau de la production qu'au niveau de la conception.

Afin d'exploiter les questions ouvertes des enquêtes de satisfaction, le client recherchait des outils d'analyse afin de :

- automatiser la classification des verbatims clients
- obtenir une vision synthétique et structurée qui fasse ressortir les points critiques du ressenti client.

Description du processus actuel / chaîne de traitement

Jusqu'à ce jour, l'affectation des verbatims clients dans les plans de classement se faisait manuellement, chaque métier possédant son propre plan de classement.

Analyse

Identification du point d'intervention TAL

La solution TAL, intervient dès la phase d'analyse des verbatims. Elle procède, à un premier niveau, à une analyse textuelle de tous les verbatims afin d'extraire de ceux-ci les informations nécessaires à la catégorisation. La solution développée intègre, à un second niveau, des technologies de text-mining :

- La classification automatique utilisée en pré-traitement explore automatiquement le contenu des verbatims.
- La catégorisation automatique permet, de classer les verbatims.

Description de la technologie TAL applicable

La solution développée met en œuvre :

- un serveur d'analyse linguistique pour procéder à l'analyse morpho-syntaxique des verbatims (tagging, lemmatisation) et en identifier le ou les thèmes abordés.
- un serveur de classification automatique pour procéder à l'analyse typologique des verbatims existants : en explorer le contenu et en proposer une cartographie sous forme de classes.

- Un serveur de catégorisation automatique de documents pour classer les verbatims suivant les plans de classement définis, après un apprentissage sur un lot de verbatims représentatif.

Description du gain qualitatif / quantitatif attendu

- D'un point de vue qualitatif :
    - Analyse plus rapide et plus précise des retours clients
    - Amélioration de la classification des verbatims
    - Mise en évidence plus rapide des points critiques exprimés dans les verbatims
- D'un point de vue quantitatif :
    - Diminution du temps de traitement des enquêtes qualité.

Déploiement et mise en œuvre

Évolution de la chaîne de traitement

Pour atteindre ces objectifs, il a été développé une application TAL qui repose sur le couplage d'une solution de clustering (organisation automatique de documents) et de catégorisation automatisée.

- Le serveur de classification a pu organiser un ensemble non structuré de verbatims en une véritable typologie des problèmes rencontrés. Il a ainsi permis d'optimiser le plan de classement initial de l'entreprise en proposant, sur la base de groupes de 300 à 5000 verbatims, 80 plans de classement comportant chacun une dizaine de catégories. Les experts métiers sont ensuite intervenus pour valider ces plans de classement.
- Le serveur de catégorisation est ensuite utilisé pour affecter automatiquement les nouveaux verbatims clients dans ces 80 plans de classement, après un apprentissage à partir d'un jeu de verbatims témoin.

Utilisation

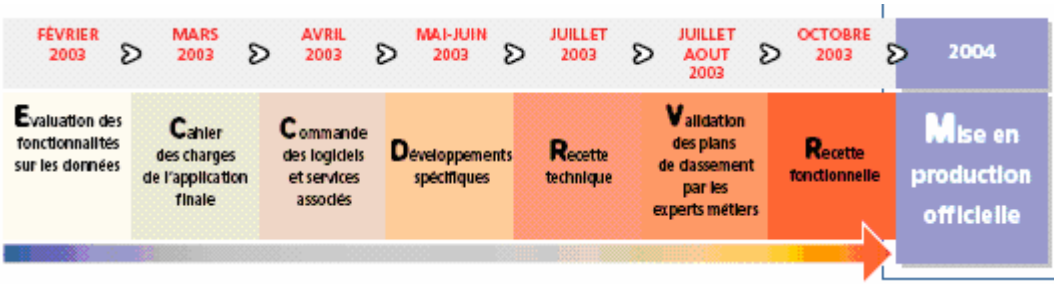
Aujourd'hui, l'application fonctionne en temps réel et concerne 400 utilisateurs des métiers de la qualité, de la conception et de la production. Les enquêtes ainsi analysées sont directement intégrées dans l'entrepôt de données du client. La simplicité d'utilisation est une des clés de la réussite de ce projet d'intégration d'outils de text-mining.

Description du traitement TAL appliqué

Analyse morpho-syntaxique, classification automatique par modèle d'apprentissage

Évaluation ROI

Coûts et délais de mise en œuvre



Gains fonctionnels / opérationnels

Nouvelles fonctionnalités fournies par l'outil :

- Existence d'un module de clustering innovant
- Classement automatique de grands volumes de verbatims clients

Gains opérationnels :

- Gain de temps : diminution du temps de traitement des enquêtes qualité. Les utilisateurs reçoivent instantanément des informations représentatives de la qualité perçue et classées par métier,
- Amélioration de la connaissance client : les métiers de la conception et de la production ont un accès direct aux ressentis clients relatifs à leur domaine respectif,
- Gain de qualité : excellente précision de la catégorisation, traduite par une grande fiabilité de l'affectation des verbatims dans les plans de classement.

Retours Utilisateurs

- Qualité de l'analyse linguistique qui permet d'atteindre une grande précision dans la catégorisation,
- Facilité d'intégration d'un vocabulaire métier spécifique à l'industrie automobile.

Gestion des brevets

Hugues de Mazancourt

Mot-clé : Information sur les brevets

Contexte, Cas Entreprise

Fournisseur d'information brevet, principalement en ligne.
Il s'agit donc d'une entreprise qui intègre le composant TAL dans son offre pour mieux répondre aux attentes de ses propres clients.

Description du besoin métier

Fournir aux clients des moyens de lire rapidement un brevet pour identifier s'il est pertinent ou non par rapport à la recherche. Une telle lecture, sans aide, est longue et coûteuse. Par ailleurs, la connaissance du contenu et de l'évolution des brevets dans son domaine est critique pour une entreprise.
Les fournisseurs d'information brevet doivent donc fournir toutes les aides possibles pour une prise de connaissance rapide de ces données.

Description du processus actuel / chaîne de traitement

L'obtention de brevets s'effectue soit par une recherche sur mots-clés, soit via la Classification Internationale des Brevets (environ 60 000 entrées). Une telle recherche sur une base de brevets nationale ou multinationale renvoie habituellement plusieurs centaines de documents. Les aspects importants (procédés innovants, champs d'application particuliers, etc.) ne se trouvent généralement pas dans le champ « Résumé » du brevet, mais quelque part de manière intentionnellement discrète, dans le corps (« texte libre») du brevet, qui fait de 10 à 50 pages.

Description des problèmes actuels

La lecture d'un lot de brevets suite à une recherche (fréquemment dans un contexte de veille technologique) est extrêmement consommatrice de temps et ne peut être confiée qu'à un expert. Par ailleurs, cette activité est aussi cruciale du point de vue de l'entreprise cliente du fournisseur d'informations brevets.
Les aides que propose le fournisseur du point de vue de la recherche ne suffisent généralement pas à réduire de manière significative le nombre de documents à lire.

Analyse

Le système va fournir une aide à la lecture en extrayant les passages essentiels d'un brevet.

Identification du point d'intervention TAL

Le point d'intervention se situe (fonctionnellement) entre la recherche et la fourniture des documents au client : ceux-ci vont être enrichis par une mise en exergue des passages-clés, en identifiant :

- l'objet précis du brevet,
- les avantages sur les inventions précédentes (aspects innovants),
- les déficits des brevets portant sur le même sujet (critiques),

- les « revendications » indépendantes du brevet

Le fournisseur d'information brevets fournit donc, en plus du texte du brevet (et de toutes les données annexes), un « résumé » présentant ces passages-clés.

Description de la technologie TAL applicable

La technologie est essentiellement une technologie de text-mining. Elle s'appuie sur la collecte de formulations récurrentes dans les brevets ainsi que sur des marqueurs lexicaux qui introduisent des éléments rhétoriques précis. Tous ces éléments sont mis en œuvre dans un système à base de règles exploitant une analyse lexicale du texte et sa structure formelle.

Description du gain qualitatif / quantitatif attendu

Le gain s'exprime à la fois en temps d'accès à l'information pertinente et en volume d'information « récupérée », c'est-à-dire l'information qui aurait été ignorée, faute de temps ou par fatigue humaine à prendre en compte tous les documents.

Déploiement et mise en œuvre

La mise en œuvre s'effectue sous la forme d'un WebService qui prend en entrée le document XML décrivant le brevet et le renvoie avec ses annotations. Ce WebService est intégré à une Interface Homme-Machine (IHM) de visualisation (fonctionnant via un butineur Internet), permettant une navigation rapide dans le texte entre les passages soulignés.

Évolution de la chaîne de traitement

Le Professionnel des Industries de la Langue fournit le Web Service qui effectue l'analyse du brevet, le fournisseur d'informations brevet l'intègre dans son offre produit via l'IHM mentionnée ci-dessus.

Description du traitement TAL appliqué

Analyse textuelle, analyse lexico-sémantique, moteur de règles basé sur la phraséologie recensée pour ce type de documents.

Évaluation ROI

Coûts et délais de mise en œuvre

Confidentiel

Gains fonctionnels / opérationnels

Gain de temps important.
A défaut d'utiliser cet outil, les clients peuvent acheter des résumés de brevets auprès de fournisseurs spécialisés qui sont coûteux (car réalisés à la main) et ne sont disponibles que plusieurs semaines, voire plusieurs mois après la publication du document initial.

Retours Utilisateurs

Retour utilisateurs très positifs. L'application est en cours d'intégration en version de production.

E-commerce

Claude de Loupy

Mots-clés : Ee-commerce, Logiciel d'analyse

Contexte, Cas Entreprise

Entreprises de vente par correspondance ayant un catalogue en ligne. Il s'agit d'entreprises ayant un catalogue assez conséquent.

Description du besoin métier

Le but de ces entreprises par rapport à leur site Internet est d'augmenter le volume des ventes. Pour cela deux axes sont privilégiés :

- permettre au client d'aller au plus vite au produit qui l'intéresse. Il s'agit donc de réduire le nombre d'opérations au maximum entre l'entrée sur le site et la fin de la commande ;
- proposer au client des produits du même type ou complémentaires pour pousser les ventes (*Push*).

Description du processus actuel / chaîne de traitement

Les catalogues en ligne sont construits sur une base de données de produits avec des informations comme le type de produit, son nom, son prix, généralement du texte libre et des informations dépendant du type de catalogue : marque, couleur, etc. Un site Internet est alors construit à partir du classement de ces produits. Mais il est généralement long de trouver ce que l'on cherche en utilisant la hiérarchie mise en place.

Description des problèmes actuels

Certains sites sont très difficiles d'accès pour des personnes n'ayant pas un minimum de connaissance de l'Internet et de l'informatique. Beaucoup de clients potentiels sont découragés par la complexité et la longueur des opérations.

Le moteur de recherche utilisé peut être très rudimentaire (ne permet pas de profiter du texte libre ou alors génère des réponses confuses).

Les liens entre les produits pour le *Push* sont faits manuellement, ce qui prend beaucoup de temps.

Analyse

Les technologies du TAL permettent une analyse plus fine des besoins du client en lui donnant la possibilité d'entrer une requête précise ou de comprendre, « intuiter » une question imprécise ou mal formulée, voire mal écrite.

Identification du point d'intervention TAL

L'analyse de la requête permet d'interroger la base de données en extrayant les caractéristiques présentées par l'utilisateur pour les transformer en requête SQL. Ainsi, il est possible d'analyser la requête "pantalon noir en velours à moins de 60 euros" pour en extraire les informations:

- article : pantalon
- couleur : noir
- matière : velours
- prix : < 60 euros

Une requête SQL est alors envoyée à la base de données ainsi qu'au moteur de recherche évolué sur le texte libre.

Description de la technologie TAL applicable

Les techniques généralement utilisées pour ce type d'analyse sont des automates dédiés. Il est donc nécessaire de reproduire des automates et des lexiques quand un nouveau type de client demande à disposer de ce module.

Par ailleurs, il est possible d'utiliser des fonctions de rapprochement entre les descriptifs des produits. Le *Push* peut ainsi être amélioré.

Description du gain qualitatif / quantitatif attendu

Le gain concerne principalement le temps mis par le client potentiel pour trouver un article qui peut l'intéresser. Il en résulte une plus grande satisfaction de ce client et surtout une diminution importante de la perte de clients potentiels (ceux qui se seraient détournés du site parce qu'ils n'auraient pas trouvé un article qui existe pourtant dans la base).

Par ailleurs, une gestion correcte des fichiers de journalisation (logs) permet de savoir quels sont les produits les plus vendus et ainsi de les mettre en valeur.

Déploiement et mise en œuvre

Évolution de la chaîne de traitement

Installation

L'installation d'un tel moteur est relativement simple pour peu qu'il puisse s'interfacer avec une base de données, être lui-même interrogé comme une base de données.

Le moteur de recherche devra être capable de supporter une charge de travail très lourde lors des interrogations, certains sites de e-commerce recevant plus de 1 million de requêtes par jour.

L'installation se fait généralement sur des PC sous Windows ou Linux. Certaines entreprises travaillent aussi sous Sun Solaris ou sous IBM/AIX.

Utilisation

L'utilisation est totalement transparente pour le client qui ne voit que le résultat des analyses. Il peut cliquer sur tel ou tel lien qui lui est proposé : visualisation du produit proposé, navigation dans des produits similaires (même type de vêtement par exemple) ou liés (un chemisier pour aller avec une jupe), validation de l'achat, etc.

Pour l'administrateur, les choses ne sont guère plus complexes. Des logs sont générés à chaque requête, permettant le suivi des opérations et une correction simple des règles d'interrogation en cas de mauvais aiguillage flagrant.

Description du traitement TAL appliqué

Les traitements de TAL sont de plusieurs natures :

- analyse morpho-syntaxique : afin d'analyser correctement la description des articles, il convient de disposer d'une bonne analyse morpho-syntaxique pour pouvoir repérer des patrons prédéfinis permettant de répondre efficacement à l'utilisateur.
- transducteurs : des suites de transducteurs (sorte d'automates) permettent d'analyser finement une requête afin d'en extraire les informations capitales qui seront transformées en requête SQL.

- synonymie/dérivation : des liens de synonymie ou de dérivation sont nécessaires dans ce type d'application (lecteur de CD portable --> balladeur CD, tenue de sport --> tenue sportive).

Évaluation ROI

Coûts et délais de mise en œuvre

Les coûts d'installation sont de quelques jours pour un catalogue de grande taille. Une machine doit être dédiée au moteur de recherche car le traitement des requêtes demande des ressources non négligeables.

Gains fonctionnels / opérationnels

Nouvelles fonctionnalités fournies par l'outil :

- Réponse plus précise aux requêtes des utilisateurs
- Transformation d'une requête utilisateur en requête SQL
- Compréhension de requêtes imprécises ou mal formulées
- Envoi d'articles similaires (*Push*)
- Statistiques sur les logs

Gains opérationnels :

Gain important dans l'utilisation du site Internet (multiplication par 4 du nombre d'utilisateurs en 3 mois) du fait de la plus grande facilité de navigation et la satisfaction plus grande des usagers.

Augmentation du chiffre d'affaires et de la satisfaction des utilisateurs qui trouvent les produits qu'ils cherchent et les achètent donc davantage (amélioration des taux de conversion et de satisfaction).

Retours Utilisateurs

Dans le cas de plusieurs catalogues en ligne, des demandes ont été faites pour d'autres langues. Ainsi, le système de recherche de catalogue a déjà été déployé, pour une même entreprise, dans 6 langues européennes et 3 langues asiatiques. D'autres langues seront déployées dans les mois qui viennent (7 langues européennes et 1 langue asiatique).

La possibilité de gérer les catalogues de manière transparente (le moteur de recherche est juste une sur-couche) et de profiter des avantages d'une analyse linguistique fine est très appréciée des entreprises de VPC.

Terminologie d Entreprise

Guillaume Mazières
Sylvie Guillemin-Lanne

Mots-clés : Terminologie d Entreprise, communication, cohérence.

Contexte, Cas Entreprise

Les grands groupes automobiles doivent produire et maintenir une documentation technique dans de nombreuses langues. Ils doivent l adapter à différents environnements réglementaires, répondre rapidement aux conditions sans cesse changeantes du marché, ainsi qu aux attentes des consommateurs.

Description du besoin métier

Aujourd hui, la production de cette documentation technique est rendue complexe du fait de la mondialisation. Pour qu elle se réalise dans les meilleures conditions, elle doit s appuyer sur un composant indispensable : une terminologie d entreprise cohérente.

Description du processus actuel / chaîne de traitement

La mise en place d une telle terminologie d entreprise a nécessité l exploration d un fonds documentaire multi-sources, contenant des millions de termes candidats.
Le processus de construction de terminologie aurait requis un budget considérable et induit des délais importants s il n avait pas été possible de l automatiser, en facilitant la construction de cette terminologie d entreprise.

Description des problèmes actuels

Quelle que soit la pièce, le véhicule ou encore le process industriel que décrit un document, il faut s assurer que le rédacteur et le lecteur auront la même compréhension de chaque terme utilisé. L utilisation d une terminologie confuse engendre des coûts élevés de traduction ainsi que des problèmes d incompréhension, qui peuvent avoir des effets négatifs sur la communication interne et externe.

Analyse

Identification du point d'intervention TAL

La solution TAL, intervient dès l exploration des fonds collectés. Le TAL permet de procéder à l analyse morpho-syntaxique des corpus et, d extraire de ceux-ci des termes candidats.

Description de la technologie TAL applicable

La solution de création de terminologie s appuie sur un moteur linguistique multilingue qui enchaîne les étapes linguistiques suivantes :

- l identification de la langue,
- la lemmatisation,
- l analyse morphologique,
- la désambiguïsation morpho-syntaxique,
- l extraction d entités.

Il est disponible en 12 langues (anglais, allemand, espagnol, français, grec, hongrois, italien, néerlandais, polonais, portugais, russe, tchèque).

Description du gain qualitatif / quantitatif attendu

- D'un point de vue qualitatif :
  - Créer une terminologie d'entreprise homogène et cohérente
  - Bénéficier d'une grande qualité de service
- D'un point de vue quantitatif :
  - Diminuer les temps de traitement,
  - Réduire l'effort manuel à son minimum

Déploiement et mise en œuvre

Il a été fourni une solution TAL de création de terminologie d'entreprise en récupérant des données existantes. La construction de la base terminologique a été réalisée à l'aide du moteur d'analyse linguistique multilingue. De grosses volumétries de sources clients ont été explorées afin de proposer des termes candidats et, partant, de construire un thésaurus multilingue cohérent.

Évolution de la chaîne de traitement

Installation / Utilisation

La solution implémentée présente une chaîne de traitement qui prend en entrée les données existantes du client quel que soit leur format d'origine (base de données, Excel, texte, etc.) et procède à une série de contrôles par des experts et de consolidations afin de réduire le nombre de termes candidats de plusieurs millions à quelques dizaines de milliers et de les présenter dans un format consolidé et vérifié.

Les différents contrôles et consolidations ont pu être effectués grâce aux technologies linguistiques telles que l'analyse morpho-syntaxique multilingue des données fournies en entrée. Les langues impliquées sont l'anglais, l'allemand, le français, l'italien, l'espagnol et le portugais.

Description du traitement TAL appliqué

Analyse morpho-syntaxique, extraction de groupes nominaux, extraction de terminologie multilingue,

Évaluation ROI

Coûts et délais de mise en œuvre

Le traitement manuel n'était d'évidence pas une solution réaliste : considérant que la vérification manuelle des différents champs associés à un terme candidat peut prendre jusqu'à 20mn ou plus par terme. Ceci doit être multiplié par le nombre de termes candidats à traiter, soit un plus de 20 moi.

Par ailleurs, l'automatisation du processus de constitution de terminologie a permis de d'affiner la méthodologie de traitement des données existantes, les règles d'extraction automatique pouvant être revues et corrigées après chaque étape, afin que le résultat final, la représentation cible des données existantes, soit conforme aux souhaits.

Gains fonctionnels / opérationnels

Nouvelles fonctionnalités fournies par l'outil :
Etiquetage des termes extraits (genre, nombre, …)

Gains opérationnels :

- Rapidité d'exécution : La solution développée peut traiter plusieurs centaines de milliers de termes par heure. Elle réalise des contrôles et des modifications qui nécessiteraient plusieurs mois d'un traitement manuel attentif.

- Amélioration de la qualité des documents et de la cohérence des traductions : cette terminologie partagée par l'ensemble du groupe assure une rédaction cohérente, indispensable à la bonne compréhension des lecteurs.

- Traduction plus rapide : L'utilisation d'une terminologie validée facilite les travaux de traduction car celle-ci accroît la fréquence de réutilisation des passages déjà traduits.

- Réduction globale des coûts et des délais d'introduction des produits sur le marché : Une terminologie d'entreprise cohérente réduit le risque de rappels de documentation liés à des problèmes de rédaction ou de traduction. Une documentation précise et rapidement disponible participe à la réduction des délais de mise sur le marché des véhicules.

Retours Utilisateurs

"Il nous est apparu évident que, vu le volume de données que nous avions à prendre en considération, un processus d'automatisation était indispensable. Ce projet nous a permis de réduire l'effort manuel à son minimum et de bénéficier d'une qualité de service incomparable. "

Gestion des candidatures

Fabienne Gire

Mots-clés : Ressources Humaines, Recrutement, Analyse des candidatures, Gestion des Compétences, E-recrutement

Contexte, Cas Entreprise

Les grandes entreprises et les grosses PME, qu'elles soient en phase active de recrutement ou non, reçoivent de gros volumes annuels de candidatures (de 20 000 à 200 000 CV) répartis selon des flux web (site de l'entreprise + job boards), mail, et courrier papier.

Description du besoin métier

Les responsables RH doivent traiter toutes ces candidatures : au minimum répondre à l'ensemble des candidats (image de marque de la société), détecter au plus tôt les profils susceptibles de répondre aux attentes de l'entreprise, contacter les candidats, les rencontrer, etc.

Description du processus actuel / chaîne de traitement

Aujourd'hui, les chargé(e)s de recrutement et leurs assistant(e)s ouvrent les enveloppes et les e-mails, lisent les CV et lettres de candidatures, saisissent les informations dans la base de données de candidats de l'entreprise, et répondent aux postulants : le travail de réception des candidatures est donc essentiellement manuel.

Description des problèmes actuels

Ce processus manuel est très coûteux en temps et en argent.
Les conséquences sont notamment :

- Délais trop longs : plusieurs semaines pour envoyer un accusé de réception.
- Perte de profils potentiellement intéressants pour l'entreprise : la plupart du temps, les RH ne gardent trace que des candidats pouvant correspondre à un profil recherché à un instant donné.
- Incomplétude des informations archivées : les chargé(e)s de recrutement n'ont souvent le temps de saisir qu'une petite partie des informations envoyées par le candidat (contact par exemple), ou bien des informations non qualifiées (CV électronique global, sur lequel on ne pourra faire que des recherches texte libre).

Analyse

Les technologies TAL permettent :

- l'automatisation du processus d'absorption des flux entrants de candidatures, en amont de leur exploitation proprement dite par un logiciel de gestion des CV. L'étape de saisie manuelle des informations est évitée
- l'optimisation de l'utilisation d'un logiciel de gestion de candidatures.

Identification du point d'intervention TAL

Le logiciel d'analyse automatique de CV intègre un module TAL d'extraction d'information qui repère les informations caractéristiques dans le texte original afin de segmenter le CV en zones, puis analyse les informations pertinentes.

Il extrait et qualifie/normalise les données concernant l'état civil, la formation, l'expérience professionnelle, les compétences du candidat...

Les résultats sont ensuite envoyés à un système expert qui génère les résultats finaux.

Description de la technologie TAL applicable

Le module d'extraction d'information s'appuie sur un moteur linguistique (détection de la langue des documents, analyse morpho-syntaxique du texte), des lexiques spécialisés ainsi que sur des règles morpho-syntaxiques et sémantiques.

Description du gain qualitatif / quantitatif attendu

- reconnaissance, qualification et stockage intelligent et systématique des informations fournies par les candidats, de façon plus exhaustive que ce que permettait le traitement manuel ;
- réduction du temps de traitement des candidatures.

Déploiement et mise en œuvre

Évolution de la chaîne de traitement

Installation

Le logiciel d'analyse de CV s'installe très facilement sur un ou plusieurs postes de travail. Il surveille automatiquement l'arrivée de nouvelles candidatures dans des répertoires (local, FTP...) ou des boîtes mail (par exemple des adresses spécialisées comme recrutement@monentreprise.com).

Utilisation

1) Les CV reçus au format électronique sont, plusieurs fois par jour, automatiquement récupérés et analysés par le logiciel.

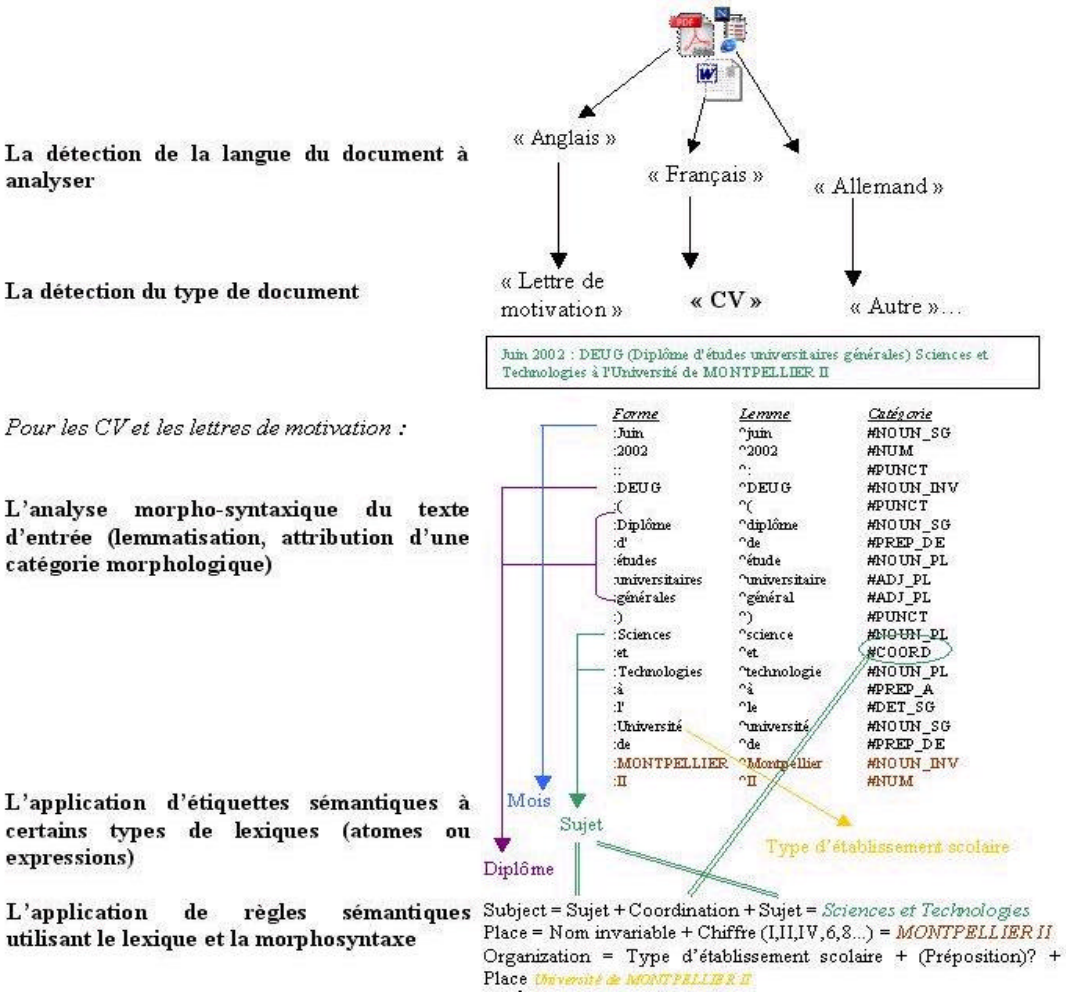
Au format papier, les CV sont, de façon quotidienne ou hebdomadaire, scannés et transmis à un logiciel d'OCR (Reconnaissance Optique de Caractères) avant d'être analysés, ce qui permet de les conserver sur un support numérique.

2) Le logiciel analyse le contenu des textes (CV, lettres de motivation) et génère des résultats (sous plusieurs formats, en particulier XML) qui sont automatiquement injectés dans une base de données de candidatures.

3) Le recruteur peut directement interroger cette dernière selon les critères de l'analyse automatique (formation, expérience professionnelle, compétences, localisation géographique...) tout en conservant la possibilité d'accéder au CV original pendant sa recherche multi-critères.

Description du traitement TAL appliqué

La technologie de Text Mining utilisée dans ce type de logiciel procède à :



Information extraites :

0 /Date *Juin 2002*
(0) 0: ending date *Juin 2002*
(0) 0: Year *2002*
1 /DegreeOrg *DEUG (Diplôme d' études universitaires générales) Sciences et Technologies à l' Université de MONTPELLIER II*
(1) 0: /Degree Subject *DEUG (Diplôme d' études universitaires générales) Sciences et Technologies*
(0) 0: /Degree *DEUG (Diplôme d' études universitaires générales)*
(0) 0: @Norm@DEUG *DEUG*
(0) 1: @Norm@DEUG *Diplôme d' études universitaires générales*
(0) 1: /Subject *Sciences et Technologies*
(1) 1: /Organization *Université de MONTPELLIER II*
(1) 0: /Place/PlaceNUM *MONTPELLIER II*
(0) 0: /Loc country.France *MONTPELLIER*

Ces traitements permettent de repérer et de normaliser les informations pertinentes avant de les renvoyer à un système expert, qui génère des résultats finaux **qualifiés** importés dans la base de données :

<DEGREE_UNIT>
<YEAR>2002</YEAR>
<DEGREE>DEUG</DEGREE>
<SUBJECT>Sciences et Technologies</SUBJECT>
<LEVEL>BAC+2</LEVEL>
<ORGANIZATION>Université de MONTPELLIER II</ORGANIZATION>
</DEGREE_UNIT>

Évaluation ROI

Coûts et délais de mise en œuvre

Les coûts d'installation d'un logiciel d'analyse automatique de CV sont minimes.

Il est préférable, mais non nécessaire, de consacrer un poste de travail à la tâche. L'installation et l'utilisation sont très simples et ne nécessitent pas de formation supérieure à une demi-journée. Il faut compter également une demi-journée pour paramétrer la surveillance automatique des flux de candidatures (obtention des autorisations, etc).

Par conséquent, les délais de mise en production ne dépassent pas une à deux journée(s), une fois le matériel de base disponible.

Gains fonctionnels / opérationnels

Nouvelles fonctionnalités fournies par l'outil :

- envoi automatique d'un accusé de réception électronique aux candidats,
- routage automatique des candidatures vers des recruteurs particuliers (en fonction du profil du postulant ou du contenu de sa lettre de motivation),
- normalisation et qualification de certaines informations : par exemple, les CV des détenteurs de DEUG, DUT, BTS, Diplôme Préparatoire aux Etudes Comptables et Financières… seront retrouvés par une recherche de candidats ayant un « Bac+2 » dans le logiciel de gestion des candidatures,
- possibilité de faire des statistiques sur les profils des candidats de l'entreprise,
- si l'entreprise souhaite conserver son système de formulaire web de candidature, génération automatique d'un formulaire à valider par le candidat à partir de l'analyse du CV déposé sur le site…
- possibilité de rendre un CV anonyme en ne fournissant au recruteur final que les informations relatives à la formation, l'expérience professionnelle et les compétences du candidat.

Gains opérationnels :

La mise en production d'un logiciel d'analyse automatique de CV montre un gain très important dans le temps de traitement des candidatures. Par exemple, pour environ 160 CV papier reçus par jour, la saisie manuelle des informations dans la base de données d'un grand compte prenait 480 minutes. Le traitement automatique prend 90 minutes.

La masse de données traitée est accrue : les entreprises peuvent conserver dans leur base de candidats la totalité des informations contenues dans les CV de tous les postulants. On augmente ainsi le ROI du logiciel de gestion de candidatures.

Retours Utilisateurs

Les DRH ayant déjà adopté un logiciel d'analyse automatique de CV sont très satisfaites du gain de temps, de la couverture accrue et des nouvelles fonctionnalités, car cela leur permet d'une part de gagner en productivité, d'autre part de consacrer plus de temps à la recherche multicritères intelligente et à la rencontre de candidats.

En outre, certaines normes qualité exigent que l'entreprise réponde à tous les candidats, que leur profil l'intéresse ou non. Le logiciel d'analyse automatique des CV permet d'effectuer cette tâche de façon systématique et dans les délais les plus brefs.

Enfin, les utilisateurs ayant à traiter des CV au format papier sont ravis de la possibilité de dématérialisation, le processus de traitement des CV étant alors intégré à la Gestion Electronique de Documents.

Dans le cas d'une grande banque française, le processus de traitement du CV papier a été entièrement réorganisé autour de l'usage quotidien d'un analyseur de CV pour un gain de productivité d'un facteur cinq.

Moteurs de recherche

Eric Debonne

Contexte, Cas Entreprise

Le moteur de recherche Intranet est devenu un centre d'intérêt, fonctionnant d'abord pour les bases documentaires, il s'est ensuite élargi lors de la mise en place d'Intranet global dans l'entreprise.

Une première phase d'équipement a consisté à remplacer des applications en technologies client / serveur ayant un accès à l'information via une hiérarchie de termes pré déterminée et via quelques mots clés à choisir dans une liste.

Le besoin actuel est de faire évoluer une application de recherche existante, car elle ne répond plus aux attentes, de nouveaux outils plus riches fonctionnellement existent et l'utilisation de l'application est devenue courante.

Les entreprises sont des grands comptes, cabinets de conseils ou banques.

Il s'agit d'applications d'accès à l'information d'entreprise caractérisée par :

- une majorité d'information interne pour un nombre conséquent d'utilisateurs (ordre de grandeur de 10 000 à 30 000)
- utilisée par un département gérant une information très ciblée d'un minimum de 10 000 documents et représentant une des principales connaissances liées à leur métier.

Un des services de quelques centaines d'avocats se base sur l'ensemble de ses expériences afin de répondre mieux et plus rapidement à ses clients dans le cadre des propositions autour de ses différentes prestations. Il s'agit d'un historique de missions de 30 000 à 50 000 documents dont 80% sont rédigés en français et le reste en anglais. Les documents sont gérés dans un environnement GED.

Le service rendu doit permettre avant tout un gain de temps afin de rendre l'activité plus productive.

Dans un autre cas, le moteur de recherche s'applique sur un environnement Intranet afin de permettre une recherche d'information sur l'ensemble des données publiées.

Description du besoin métier

Une première application de recherche a été mise en place depuis 2 à 3 ans. Les éléments déclencheurs d'évolutions sont les suivants :

- Une maintenance difficile,
- Une architecture technique complexe, qui n'a pas forcément évolué,
- Une application qui n'a pas été suivie et qui devient obsolète,
- Une application qui ne correspond plus aux besoins des utilisateurs.

A partir de ce constat, une étude de besoins est effectuée en parallèle de la question qui se pose : peut-on conserver les solutions logicielles acquises et est-il pertinent de les comparer avec d'autres ?

Voici quelques uns des besoins les plus souvent exprimés :

- recherche d'une architecture technique simple,
- la syntaxe de requête doit être libre,
- la gestion des dérivés simples est obligatoire,
- le référentiel terminologique existant doit être ré-utilisé,
- la qualité des résultats doit être au moins équivalente ou supérieure selon les cas,
- le logiciel doit prendre en compte les documents anglais pour une recherche en anglais et donner les moyens d'une qualité de résultats au moins équivalente à l'existant en français,
- la qualité des résultats est aussi importante que la convivialité et les fonctionnalités de l'application permettant un accès facile à l'information,

- le suivi qualité de l'application devient un point clef pour la pérennisation de l'investissement dans le temps.

Description du processus actuel / chaîne de traitement

Les documents sont gérés soit au sein d'un logiciel de GED (par exemple Hummingbird DM, Lotus Notes, Documentum), soit au sein d'une base de données, soit sur disque.

Les fichiers sont indexés par un moteur de recherche accédant aux données.

La recherche est plus ou moins pertinente selon les solutions choisies : chaîne de caractères, mots clés, algorithmes statistiques et/ou linguistiques. La visualisation des documents se fait régulièrement après une conversion en HTML, sinon avec une reprise du format natif. La mise en surbrillance des termes concernés par la recherche est active.

Description des problèmes actuels

Absence de suivi de l'application.

L'application installée n'a pas évolué dans la plupart des cas. Il s'ensuit des problèmes de type :

- maintenance d'une architecture obsolète,
- pas de compétences pour un paramétrage, une correction ou une évolution,
- contour de l'application ne répondant plus aux besoins de l'utilisateur : mise à jour des bases documentaires pertinentes, ergonomie,
- qualité de réponse dépréciée : le référentiel terminologique n'est plus à jour, le paramétrage n'est plus adéquat,
- pas de prise en compte des retours des utilisateurs,
- pas de communication d'utilisations.

L'application ne permet pas un accès facile et rapide à l'information :

- moteur d'indexation et de recherche par chaîne de caractères ou mots clés,
- pas d'utilisation de référentiel terminologique : synonyme, extension des sigles,
- pas de prise en compte de pluriels, de mots composés ...
- manque de fonctionnalité d'aide à la lecture de la liste de résultats : résumé, extrait,
- manque d'aide à l'affinage ou reformulation de la requête,
- manque d'outils de structuration de la liste de résultats : sélection par attributs : type, date, format, auteur, origine….; catégorisation, clusterisation
- manque d'outils de gestion des informations recherchées et retrouvées (activité post recherche) : requêtes, paniers de documents retrouvés ...

Analyse

Identification du point d'intervention TAL

Voici quelques points sur lesquels le TAL peut s'appliquer :

- suivi qualité : identification des expressions et des termes les plus souvent utilisés ou nouvellement utilisés lors de la recherche et identification de termes proches dans le corpus de document ;
- création automatique et utilisation d'un référentiel terminologique ;
- analyse linguistique de la requête et des documents lors de l'indexation ;
- complément de qualité par l'analyse de la pertinence des résultats avant l'affichage ;
- suivi qualité et aide à la lecture de la liste de résultats : extraction de termes : noms de personnes, sociétés, chiffre, action du domaine (achats, ventes par exemple) ...
- analyse des retours des utilisateurs : demandes fréquentes, éléments d'indice de satisfaction ;
- structuration de la liste de résultats : clusterisation, catégorisation automatique ;
- ajout d'un résumé ou d'un extrait des documents dans la liste de résultats ;
- correction des fautes de la requête.

Description du traitement TAL appliqué

Les technologies TAL peuvent être utilisées seules mais sont souvent couplées à des algorithmes statistiques :

- analyseurs linguistiques de requêtes,
- extraction linguistique de contenu de document : suivi qualité, résumé, analyse de la liste de résultats, indexation linguistique,
- Text Mining pour la création automatique d'un référentiel terminologique, extraction de données spécifiques : nom de personnes, de sociétés …, clusterisation, catégorisation.

Description du gain qualitatif / quantitatif attendu

Le gain qualitatif attendu lors de l'utilisation de ces technologies est un accès plus simple et rapide à l'information : l'application présente des résultats plus précis et est capable de trier, structurer, expliciter les résultats afin de visualiser les différentes informations disponibles dans le contexte de l'utilisateur.

Déploiement et mise en oeuvre

Dans le contexte d'entreprises visées, la technologie TAL n'est pas une valeur ajoutée en tant que technologie mais pour le service rendu dans l'environnement de recherche. Les composants TAL sont donc soit une base du logiciel mis en œuvre, soit utilisés par des fonctionnalités du logiciel.

Le plus couramment, seule la gestion du référentiel terminologique est visible par l'équipe projet client car cette fonctionnalité nécessite des interventions de documentalistes ou d'experts métiers de l'entreprise lors de la mise en œuvre et du suivi de l'application.

Évaluation ROI

Coûts et délais de mise en œuvre

Si l'on considère les fonctionnalités utilisant une technologie TAL, le calcul du ROI correspondant à l'apport de cette technologie n'est pas souvent visualisable. En effet, pour une application de recherche globale, il s'agit de l'existence ou non d'un outil d'accès à l'information.

Pour les fonctionnalités directement issues du TAL, il s'agit de gain de temps de recherche, d'exhaustivité, de complétude et de précision de l'information retrouvée. Ces apports sont mesurables dans des environnements où l'utilisateur dispose d'un temps compté et facturé, comme ce peut être le cas d'un avocat. Mais dans d'autres environnements, il faut mesurer le gain de temps obtenu par utilisateur et estimer le coût de ce temps, ce qui est particulièrement complexe.

Cependant, l'étude initiale du besoin lié à l'application doit révéler le ou les services répondants aux plus fortes valeurs ajoutées de l'application dans son contexte d'utilisation. Ces services sont plus facilement mesurables par le responsable du projet client.

La mise en œuvre doit prendre en compte la charge de travail en amont, s'il y a constitution d'un référentiel terminologique, la charge lors de la mise en œuvre et la maintenance de ce référentiel. Ces tâches peuvent s'effectuer par un service externe.

Différents niveau d'investissements peuvent être effectués selon la particularité du projet : d'une approche très automatisée à une approche très personnalisée par une compétence humaine. Ce choix s'effectue selon les besoins à forte valeur ajoutée identifiés, le public concerné (experts ou non), les documents concernés (métier précis ou Intranet global) et l'apport d'un pourcentage de qualité supplémentaire.

La charge de suivi qualité globale (dont les aspects TAL) de l'application est un des centre d'intérêt des entreprises concernées.

Retours Utilisateurs

Les retours utilisateurs sont de plus liés à l'utilisation et la fréquence d'utilisation de l'application.

Les standards

Les applications des industries de la langue interagissent avec l'environnement informatique qui les accueille, et, de plus en plus, interagissent entre elles. La question des standards de représentation des données consommées ou produites dans le cadre des applications traitant de l'Information non-structurée a donc une importance particulière. Ainsi, on doit pouvoir garantir l'échange et l'interopérabilité :
- des données textuelles,
- des ressources linguistiques et terminologiques
- des méta-données.

La première partie de cette section traite des données textuelles et des ressources. La deuxième partie de cette section traite des méta-données et particulièrement du Web sémantique.

Standards des données textuelles et des ressources

Gil Francopoulo

Alors que les standards ("de facto" ou "de jure") sont d'un usage répandu depuis longtemps dans les infrastructures informatiques, on observe leur diffusion timide dans les applications TAL.

Le besoin d'utiliser des standards existe, et il concerne :
- l'interopérabilité des différents composants au sein d'une application fondée sur les TAL ;
- l'interopérabilité de l'application fondée sur le TAL vis-à-vis des autres applications informatiques qu'elles soient locales ou distantes ;
- la gestion cohérente d'environnements multilingues complexes. Citons simplement deux exemples de difficulté : l'Europe comporte maintenant plus de 20 langues très différentes les unes des autres. Quel rapport existe-t-il entre le maltais qui est une langue sémitique (arabe pour faire simple) et l'estonien qui est une langue finno-ougrienne ? L'autre exemple est le chinois qui autorise plusieurs trans-litérations du même mot comme « non-spaced pinyin » par opposition à « spaced pinyin and tone ».

Aucun acteur n'ayant réussi en 30 ans à imposer un standard "de facto", la seule solution consiste à définir un ou plusieurs standards "de jure".

La direction qui a été adoptée récemment à l'ISO consiste à définir une famille de normes destinées au TAL au sein de l'ISO-TC37. Ces normes opèrent à deux niveaux.

Les normes de bas niveau traitent des valeurs constantes et ne sont que faiblement structurées. Ce sont les normes existantes depuis longtemps pour les codes de langues (ISO-639), les codes de scripts (ISO-15924), les codes des pays (ISO-3166) et Unicode pour le codage des caractères. Ces normes sont en train d'être complétées par une norme spécifique aux constantes linguistiques afin de fixer des valeurs comme /feminine/ et de statuer que /grammatical gender/ se définit par /masculine/ et /feminine/ dans la langue française, alors qu'il faut ajouter /neuter/ pour la langue allemande par exemple. Les constantes en question portent sur la morphologie (comme dans l'exemple) mais aussi sur la syntaxe, la sémantique et l'administration. Elles sont gérées dans un répertoire de catégories de données (Data Category Registry) dans le cadre des travaux de révision de l'ISO-12620.

Les normes de haut niveau se fondent sur ces normes de bas niveau mais sont en revanche beaucoup plus structurées. Ce sont TMF, LMF et MAF. La première (i.e. TMF pour Terminological Markup Framework ISO-16642) traite des terminologies d'entreprise qu'elles soient monolingues ou multilingues. La seconde (i.e. LMF pour Lexical Markup Framework ISO-24613) couvre les dictionnaires dans une large mesure, que ce soient les lexiques destinés au traitement automatique du langage ou bien les bases de données éditoriales servant de support de traduction aux grandes administrations. La troisième norme (i.e. MAF pour Morpho-syntactic Annotation Framework ISO-24611) traite de l'annotation des corpus que celle-ci soit effectuée par un être humain ou bien par un programme.

Dans la mesure où ces trois normes de haut niveau échangent les même constantes et qu'elles sont définies en XML, l'interopérabilité entre elles est très forte.

Les normes de haut niveau sont proches des utilisateurs puisqu'elles enregistrent les pratiques des gens du métier, alors que les normes de bas niveau sont plus du domaine de la « tuyauterie ». Mais les unes ne vont pas sans les autres.

Pour finir, notons que ces normes sont définies par des experts mandatés par leur délégation nationale respective avec une forte implication des pays des continents asiatiques, américains et européens, à raison d'un tiers chacun. En revanche, on peut déplorer que les pays à langue sémitique et africaine (Afrique du Sud exceptée) ne sont pas très impliqués. Ces pays ont évidemment d'autres urgences, on peut le comprendre. Leurs langues n'en sont pas pour autant oubliées : elles sont simplement prises en compte par des experts des autres pays. Et la représentation au sein de l'ISO n'est ni plus ni moins que le reflet du dynamisme (ou de l'attentisme) des différentes nations dans les mécanismes numériques liés à l'information, que ce soient pour des raisons économiques ou politiques.

Le Web sémantique : principes, applications et perspectives

Bruno Menon
(ce texte est adapté d'un article paru dans
Documentaliste – Sciences de l'information,
Vol. 40, N° 6)

Principes

Le projet du Web Sémantique naît de critiques bien connues adressées au réseau Internet (et, partant, intranet et extranet) sous sa forme actuelle : HTML donne des liens sans sémantique, tissant certes un réseau hypertextuel dense, mais où l'on manque de repères ; les moteurs de recherche laissent beaucoup d'opérations à la charge des internautes et leurs résultats sont souvent hasardeux ; les métadonnées sont limitées dans leur usage comme dans leur portée, peu fiables, peu utilisables et peu utilisées par les moteurs de recherche. Bref, alors que par leur volume et leur diversité, les ressources du Web sont de moins en moins exploitables sans l'aide de logiciels aux fonctions avancées, elles sont à cause de ces faiblesses peu propices aux traitements automatisés.

On entend donc mettre en place un dispositif permettant de structurer les informations du Web de façon à les rendre manipulables et "compréhensibles" par des agents logiciels. Leur visée : faciliter l'utilisation des informations et services du Web en libérant l'internaute d'une partie de sa charge cognitive, et donner l'impression d'un système homogène et cohérent en mobilisant automatiquement et de manière transparente les multiples ressources, sites et services nécessaires à l'accomplissement d'une tâche.

Mais si la vision ultime est celle d'un tout dont l'efficacité serait supérieure à celle de la somme de ses parties, le vocable "Web Sémantique" recouvre en réalité une grande variété de fonctions, dont certaines restent d'ailleurs à imaginer. En voici quelques-unes.

- La recherche généraliste bien sûr, avec le moteur de recherche sémantique : doté de capacités de raisonnement, il s'appuie sur la description formalisée et la mise en relation des différentes sources d'information pour traiter intelligemment les requêtes et présenter en une seule étape des résultats complets.
- L'exploitation et la combinaison de ressources pour accomplir une tâche spécialisée : des outils dédiés associent dialogue avec des sources hétérogènes, description des préférences des utilisateurs et raisonnement basé sur des connaissances métier pour synthétiser l'information requise.
- L'offre de services Web plus complets, avec des outils qui identifient, activent et combinent différents services pour mener à bien des opérations plus ou moins complexes de la vie quotidienne ou professionnelle, comme l'organisation d'un voyage, la souscription d'un contrat d'assurance, etc.
- La navigation sémantique, qui profite de la sémantisation des hyperliens pour orienter l'internaute dans son parcours du réseau.

Ces différents types de systèmes seront en outre, au moins dans un premier temps, bâtis autour de communautés d'intérêts, dans des domaines bien circonscrits, et pour des portions du Web, publiques ou privées : un "Web Sémantique d'entreprise", par exemple.

Il serait donc inexact de voir dans le Web Sémantique une entreprise monolithique : on parlera de Webs sémantiques, au pluriel, dès lors que des sites intègrent une ou plusieurs fonctions avancées mettant en jeu les concepts du Web Sémantique, au singulier. Car ce qui fait l'unité du projet est une communauté de principes et de méthodes, une démarche.

Les métadonnées

L'annotation des ressources du Web par les métadonnées, tout d'abord. La notion de métadonnée n'est pas nouvelle. Mais il va sans dire que nous sommes loin de l'usage plus ou moins anarchique des balises META de HTML ; on s'éloigne même quelque peu du concept élaboré dans le cadre des bibliothèques virtuelles et du Dublin Core. Bien que cette filiation ne soit pas désavouée, le rôle central qu'on entend faire jouer aux métadonnées dans le Web Sémantique laisse supposer que leur portée sera amplifiée par rapport à une approche "catalogage et indexation".

De ce point de vue, la définition qui en a été proposée, "Information associée à une ressource du Web, permettant d'en favoriser l'utilisation par un agent humain, du fait de son exploitation par un agent logiciel", est révélatrice. Assez large, elle met l'accent sur la finalité des métadonnées, sans vraiment insister sur leur nature descriptive. C'est qu'il y a à ce sujet une ambiguïté : s'agit-il de décrire des ressources numériques ou plutôt de programmer leur utilisation par des logiciels ? En réalité, les fonctions des métadonnées dans le Web Sémantique dépassent les dimensions signalétique et thématique qu'on leur connaissait jusqu'à présent. Selon le contexte et les applications, elles servent aussi de support à la gestion des droits, au recueil d'annotations diverses telles que commentaires et recommandations, à la qualification des hyperliens, à la définition de parcours de lecture ou d'assemblage de documents à la carte, etc.

Les ontologies

Pour être susceptibles d'être exploitées automatiquement, les métadonnées doivent être entièrement explicites, c'est-à-dire suivre un modèle et être exprimées dans un vocabulaire clairement et formellement définis. Les ontologies, deuxième pilier du Web Sémantique, sont le réceptacle de ces définitions. Elles modélisent les connaissances nécessaires à la description – et au traitement – d'un ensemble de ressources. On y représente les valeurs que l'on peut donner aux métadonnées et l'interprétation que les systèmes peuvent en faire, c'est-à-dire les concepts d'un domaine, les relations qu'ils entretiennent et la sémantique de ces relations, mais aussi les règles de raisonnement qui leur sont applicables.

On soulève souvent la question de l'analogie avec les thésaurus : la structuration des concepts en réseau et la normalisation de leur expression sont des points communs indéniables, mais ne doivent pas masquer les spécificités de chacun de ces instruments. Bien sûr, il est possible, et même souhaitable, que l'on tire parti de l'existant et que les thésaurus servent de point de départ à la construction d'ontologies pour le Web Sémantique. Il est toutefois probable qu'ils seront remaniés et étoffés. Par exemple, il est souvent nécessaire d'intégrer aux ontologies des connaissances sur des personnes ou des lieux, pour lesquels d'autres informations que celles portées par les relations classiques des thésaurus sont nécessaires. Ces spécificités dérivent de vocations dissemblables : les thésaurus sont adaptés à leur rôle d'outils de médiation documentaire, les ontologies doivent servir à la représentation de multiples aspects des ressources numériques ; les thésaurus sont destinés avant tout à un usage humain, les ontologies davantage orientées vers un usage par les machines (même si au cours de leur cycle de vie, les consultations humaines sont nécessaires et fréquentes).

En conséquence, les normes pour les thésaurus fixent la liste des relations utilisables et la forme des termes, mais laissent un certains souplesse dans les formats et les présentations utilisés ; pour les ontologies, on a en revanche une normalisation très stricte des formats, mais une grande ouverture dans la définition des relations nécessaires aux applications visées et dans le type de termes qui y figurent.

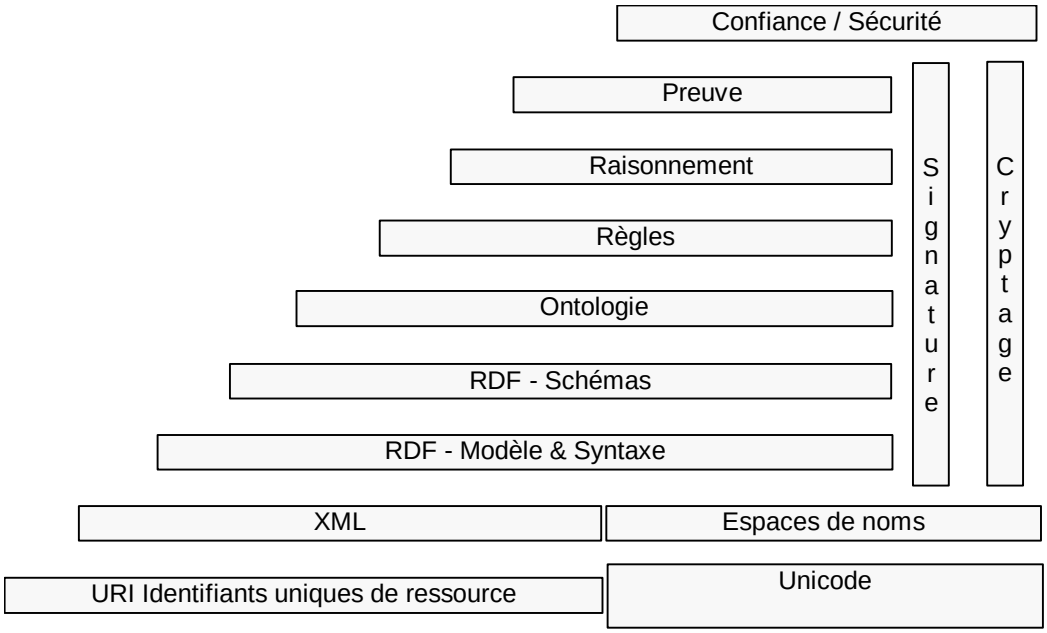
Le raisonnement

Enfin, troisième ensemble de composantes, les méthodes de raisonnement : techniques de déduction et de preuve sont indispensables à la fois pour effectuer les enchaînements impliqués par les règles d'utilisation des concepts des ontologies et expliquer le cas échéant les résultats fournis de façon convaincre l'internaute de leur validité. Le dispositif est complété par des systèmes de cryptage et de certification, qui ne sont pas spécifiques de la démarche du Web Sémantique, mais y ont néanmoins leur place, en contribuant à instaurer une confiance que le Web actuel n'inspire pas toujours.

Modèles et standards

On aura reconnu dans ce qui précède une approche similaire à celle de l'intelligence artificielle ; des bases de faits – les méta-données – sont interprétées grâce à des bases de connaissances – les ontologies – par des moteurs d'inférence, qui accomplissent ainsi des tâches complexes en simulant le comportement d'intervenants humains. Les techniques sont analogues, mais se distinguent dans le cas du Web Sémantique par leur contexte de fonctionnement et par l'échelle à laquelle elles doivent s'appliquer, inimaginable il y a vingt ans. De ce fait, l'intégration d'informations provenant de sources hétérogènes et la recherche de l'interopérabilité de sites et de systèmes implantés en divers endroits par différents acteurs revêtent une importance cruciale. D'où une volonté marquée de standardisation, problématique qui occupe une place prépondérante dans les travaux actuels.

Standardisation d'abord des formats d'encodage des informations, des méta-données et des ontologies : par son expressivité et sa souplesse, la syntaxe XML est appelée à servir de lingua franca au Web Sémantique. Standardisation également des modèles et langages permettant de décrire de façon entièrement explicite les sources d'informations et les services disponibles, mais aussi de coder dans des ontologies les connaissances nécessaires à ces descriptions. RDF (Resource Description Framework) et OWL (Ontology Web Language) seront sans doute les principaux vecteurs de ce qui constituera le noyau à proprement parler sémantique du Web. Les formalismes de raisonnement et les techniques de preuves semblent pour l'instant moins consensuels.



Le modèle en couches (*layer cake*) du Web sémantique

Applications

Dans l'entreprise

L'exemple des Webs sémantiques d'entreprise illustre l'application de la démarche du Web Sémantique à des contextes plus restreints que le Web "grand public" : les multiples problématiques de gestion des connaissances autour de ressources informationnelles qui sont, dans leur diversité, un modèle réduit de ce que l'on rencontre sur le réseau en font un champ d'expérimentation privilégié. L'échelle relativement réduite et la présence d'une culture d'entreprise rendent plus aisées la création des ontologies et la définition des usages et des utilisateurs. Par rapport à un traitement plus classique de ces problématiques, l'approche du Web Sémantique offre une garantie de stabilité et de compatibilité, du fait de l'utilisation de langages et d'outils standard pour la construction des ontologies et le recueil des annotations. Parmi les systèmes visés figurent des applications de mémoire d'entreprise distribuée, de mémoire de projet avec retour d'expérience, de cartographie de compétences pour l'aide à la recherche de partenaires commerciaux et industriels.

Portails touristiques

Pour preuve de la maturité atteinte par les techniques du Web Sémantique, et de l'intérêt qu'il y a à les appliquer à des secteurs fortement demandeurs, on peut évoquer des applications dans le domaine du tourisme, qui est l'industrie la plus numérisée. Les collectivités locales, en particulier, ont beaucoup à gagner, économiquement et en notoriété, en mettant en valeur sur le réseau leurs atouts touristiques. D'où la conception d'un service Web chargé d'organiser et de présenter les ressources documentaires d'une région, issues de sources d'information variées, locales ou non. Ce service réutilise le thésaurus de l'OMT (Organisation Mondiale du Tourisme) et en fait une partie de l'ontologie de l'application, tout en le complétant par des connaissances sur les lieux, les personnes et les objets touristiques pertinents (hébergement, patrimoine, transports).

Presse et médias

Une autre problématique sectorielle, expérimentée lors de l'Exposition Mondiale 2000 à Hanovre, concerne la presse et les médias. Le problème posé par le traitement intelligent de l'information dans ce secteur est qu'il n'est guère envisageable de décrire ou d'indexer l'intégralité de sa production documentaire : volumes importants, très rapide renouvellement et durée de vie très brève de l'information sont des contre-indications du traitement documentaire classique, sur le plan économique comme sur celui de l'efficacité. Il est en revanche possible de capturer à la fois l'univers référentiel, la sémantique et la phraséologie de la presse dans une ontologie, laquelle peut être exploitée pour rechercher intelligemment dans le texte intégral. L'approche défendue est donc celle, un peu paradoxale, d'une application de Web Sémantique sans méta-données. L'idée de déplacer l'essentiel de la charge de travail vers la formalisation des connaissances donne un aperçu de la manière dont pourraient évoluer les missions des professionnels de l'information dans l'avenir.

Perspectives

Est-ce à dire que nous touchons à l'âge d'or de l'information sur les réseaux ? Pas encore : un certain nombre de problèmes à la fois méthodologiques, techniques et organisationnels demeurent, et appellent à poursuivre les efforts.

Par exemple, sur le Web, la notion de document est plus ou moins co-extensive à celle de page, ce qui dans beaucoup de cas n'est guère satisfaisant. On sait que tout traitement documentaire suppose la délimitation de l'unité documentaire à traiter, et il n'en ira pas différemment pour le recueil de méta-données. Il importe donc de réfléchir à cette question dans le cas du document numérique sur le réseau.

Mais il faudra surtout, lors du passage à la pratique, en grandeur réelle, répondre au double défi que représentent la création et la mise en oeuvre des ontologies et la constitution des méta-données.

Les ontologies sont en théorie plus complètes, plus détaillées et plus complexes que les thésaurus, et risquent de se révéler encore plus ardues à confectionner et à maintenir. C'est pourquoi des éléments de méthodologie sont indispensables, et commencent à voir le jour. Bien que présentées initialement comme des instruments idéalement formels et raffinés, les ontologies "réelles" sont plus pragmatiquement le résultat de multiples compromis entre fonctionnalité et complexité. Il s'agit donc d'adapter leur niveau de détail à leurs visées opérationnelles, de concilier volume de concepts à représenter avec le maintien de la cohérence, et de garantir le degré de formalisation requis par les capacités des logiciels qui devront les utiliser. Au delà de la mise au point de méthodes pour créer les ontologies, l'ingénierie ontologique doit aussi se préoccuper de divers aspects relatifs à leur exploitation : ainsi, assurer l'interopérabilité de ressources hétérogènes nécessitera de fusionner différentes ontologies du même domaine ou de domaines connexes, en tenant compte de la multiplicité des points de vue possibles sur une même réalité.

Le deuxième défi du passage à la pratique réside dans l'organisation du recueil des méta-données. Qui les constitue et à quel moment sont-elles créées ? Dans le modèle documentaire traditionnel, elles le sont a posteriori, par des experts de la description des contenus ; les promoteurs du Web Sémantique voient plutôt les méta-données comme sous-produit de l'activité de production d'information. Cette dernière hypothèse est viable tant que l'on se limite à un noyau conventionnel de méta-données, de type Dublin Core, ou dans le cas d'une information produite par l'institution qui l'utilisera ; elle est moins crédible dès lors que les usages de cette information ne sont pas connus à l'avance, et des scénarios intermédiaires de partage des tâches devront être imaginés. Quel que soit le scénario retenu, il est impératif de disposer d'outils de productivité, garantissant la cohérence syntaxique (les méta-données doivent respecter le formalisme voulu), et la pertinence sémantique (les valeurs choisies doivent figurer dans l'ontologie de référence et bien caractériser les objets décrits). Certains de ces outils existent déjà, mais demandent sans doute à être perfectionnés.

Il ne fait guère de doute que les technologies de TAL, adaptées à ces diverses problématiques, pourront s'avérer un facteur crucial de réussite au moment de la mise en place d'applications de Web sémantique. Analyse discursive pour la découverte des structures et articulations logiques des documents, extraction de connaissances pour aider à constituer les ontologies, indexation intelligente pour automatiser la création des métadonnées : autant de domaines où les industries de la langue peuvent contribuer à la viabilité technique et économique du concept de Web sémantique.

Conclusion

Le fait que l'on puisse aujourd'hui parler des industries de la langue est la marque de leur maturité. Tirant ses sources des premières tentatives de formalisation de l'étude de la langue datant du début du XXème siècle, le traitement automatique des langues a évolué et continue d'évoluer, tant du côté de la recherche fondamentale ou appliquée, que de celui de sa réalisation industrielle.

Dans le domaine de la recherche, le traitement automatique des langues possède ses problématiques propres, ses écoles, ses procédures d'évaluation. Dans le domaine industriel, les industries de la langue fournissent des applications mises en œuvre par des clients qui en tirent profit et en mesurent les retours sur investissement.

On constate également que les technologies du traitement automatique des langues se banalisent et s'intègrent dans des applications et des produits grand public à mesure que la société de l'information avance.

Issues des laboratoires, allant au-delà même de leurs domaines d'application, les industries de la langue se sont adaptées aux standards existants autour de l'information et ont également contribué à les définir.

Ainsi, le Web Sémantique souligne la nécessité de prendre en compte l'organisation liée au contenu des documents. Ce projet ambitieux visant à l'interopérabilité des contenus, comme Internet visait à garantir la diffusion des documents, représente une nouvelle opportunité pour les industries de la langue, qui en sont partie intégrante.

La multi-culturalité et le multilinguisme des personnes, des entreprises et de leurs échanges, la multiplication des langues sur Internet, et la part décroissante de l'anglais dans les documents disponibles, rendent encore davantage nécessaire le recours aux industries de la langue.

Si, comme le rappelle l'introduction, les industries de la lange ont aujourd'hui l'âge de raison, elles entrent désormais dans leur maturité et comme le dit Bill Waterson : « L'enfance est courte, la maturité infinie. »

Contacts

APIL
www.apil.asso.fr
bureau@apil.asso.fr

GFII
www.gfii.asso.fr
gfii@gfii.asso.fr

Les Auteurs

Alain Couillault :

Docteur en Sciences du Langage, mention Linguistique et Informatique de l'Université de Clermont-Ferrand. Il a été successivement Directeur de Projets et Directeur de Produits chez LexiQuest et Albert Inc SA, sociétés éditrices de logiciels utilisant les technologies linguistiques. Il est co-fondateur, premier président et président d'honneur de l'APIL. Il est également co-fondateur d'Isade, société de conseil en gestion de l'Information (www.isade.com).

Eric Debonne :

Titulaire du DESS Informatique et Intelligence Artificielle de Luminy, il a été Responsable Service et Directeur avant vente Europe LexiQuest, Solution linguistique d'accès à l'Information, et consultant méthodologie Objet durant 3 ans. Il est actuellement Consultant formateur Indépendant (société Solaci). Expert moteur de recherche Intranet et Veille Internet, il a 10 ans d'expérience dans le conseil et l'accompagnement de projets d'entreprises.

Gil Francopoulo :

Docteur en mathématiques appliquées de l'Université Paris VI (Jussieu). Il a travaillé une vingtaine d'années dans le TAL chez divers éditeurs. Actuellement, il est directeur de Tagmatica. Après avoir été dans les années 90 l'un des auteurs du modèle de dictionnaire GENELEX, il est maintenant l'éditeur international de la norme ISO dédiée aux dictionnaires électroniques. De plus, il anime la gestion du registre de catégories de données de la morphologie et syntaxe dans le cadre de l'ISO-12620.

Alain Garnier :

Ingénieur diplômé de l'IIE, il finalise au sein d'Erli (SPSS-Lexiquest) ses travaux de recherche sur « les systèmes sémantiques d'extraction automatique ». Il rejoint ensuite EDS à Dallas, puis Madicia. Il fonde Arisem en 1996, éditeur de logiciel spécialisé dans le traitement de l'information sémantique où il occupe depuis lors le poste de CTO.

Fabienne Gire :

Diplômée en anglais et en Sciences du Langage, elle est titulaire du DESS ILSI (Ingénierie de la Langue et Société de l'Information) de Paris IV Sorbonne et du DEA Linguistique, Logique et Informatique de l'Université de Clermont Ferrand. Elle a travaillé plusieurs années dans la formation avant d'intégrer le monde du TAL chez un éditeur de logiciels Text Mining, et occupe actuellement le poste de consultante linguiste (responsable du département linguistique) chez KOLTECH, éditeur de Solutions pour les Ressources Humaines.

Sylvie Guillemin-Lanne :

Licenciée en russe (INALCO), elle a obtenu à Paris VII une maîtrise en linguistique & informatique, puis un DEA de linguistique formelle. Elle a travaillé successivement chez IBM France (Centre scientifique et Département des logiciels de communication en français), puis chez IBM US, au Centre de recherche TJ Watson. Actuellement, elle est Project Manager au sein de TEMIS, éditeur de logiciels en text mining, où elle assume la fonction de chef de projet clients.

Par ailleurs, elle intervient dans différentes cursus de DESS (Institut des Sciences Humaines Appliquées à la Sorbonne, Université de Poitiers) ou Mastère (Faculté d'histoire de la Sorbonne Nouvelle).

Vice-présidente de l'APIL.

Claude de Loupy :

Ingénieur et docteur en informatique. Spécialisé dans les moteurs de recherche, l'utilisation de connaissances et de traitements linguistique en gestion de l'information.

Il a été successivement Ingénieur d'études au CNRS (dans le cadre du projet européen MulText), Ingénieur de Recherche chez Bertin Technologies, puis ATER à l'Université d'Avignon et des Pays de Vaucluse, rattaché au Laboratoire Informatique d'Avignon.

Il est actuellement Responsable Recherche chez Sinequa et Maître de Conférence Associé à l Université de Paris 10.

Hugues de Mazancourt :

Président de l'APIL depuis Avril 2004. Directeur technique et co-fondateur de Lingway en 2001. 39 ans, ingénieur ENSTA, DEA Paris 1989. Expert en Traitement Automatique du Langage Naturel et développement logiciel. A ERLI, puis LexiQuest et maintenant à Lingway, il a coordonné les équipes de développement de parsers, de dictionnaires et de grammaires et a participé à la majorité des méthodes et moteurs linguistiques développés à l'occasion de grands projets comme Genelex, Graal ou le 3611.

Guillaume Mazieres :

Vice-président pour les Ventes et le Marketing de TEMIS, éditeur européen de technologies de Text Mining, qu'il a rejoint en 2001 après 8 ans à l étranger chez le fabricant de solutions de stockage de données LaCie. Après des débuts à Londres en tant qu Account Manager sur le marché des grands comptes Anglais pendant 2 ans, Guillaume Mazieres prend la direction commerciale puis la direction générale de LaCie Espagne, à Madrid. Il est ensuite nommé à la tête de la filiale canadienne à Toronto, qu il dirigera pendant 2 exercices, avant de rejoindre le siège social USA à Portland, Oregon, au poste de Vice-Président des Ventes et du Marketing pour l Amérique du Nord.

Bruno Menon :

Diplômé en Lettres et en Sciences du Langage, et titulaire du DESS de Sciences de l information et de la documentation (Institut d études politiques de Paris). Après un passage au centre de documentation contemporaine de Sciences-Po, il rejoint ERLI / LexiQuest, où il contribue à la conception des applications comme des outils et ressources linguistiques. Il a également participé à plusieurs grands projets européens (Genelex, Graal, Transterm). Il intervient dans différents cursus de formation initiale ou continue en TAL et en Gestion des connaissances (Sciences-Po, Poitiers, Paris X, ADBS).

Quelques acteurs du domaine et sources d information

AFCP : Association Francophone de Communication Parlée
www.afcp-parole.org

AFNOR : Association Française de NORmalisation
www.afnor.fr

ATALA : Association pour le Traitement Automatique des Langues
www.atala.org

ELRA : European Language Resources association
www. elra.info

Le portail Technolangue :
www.technolangue.net

Annexe A – Quelques sociétés du domaine du traitement de la langue en France[7]

4DConcept
ABILDOC
ACETIC
ALBERT
ALMAS INGENIERIE
ALOGIC
ANGLO-INSTITUTE
APPLIED SEMANTICS
ARISEM
ASCOLA
ATLANTIC INTELLIGENCE
ATOMZ
AURALOG
AUTONOMY
BABELING
BILINGUA Ingénierie Linguistique
BOWNE GLOBAL SOLUTIONS
CAP GEMINI INNOVATION
CONNEXOR
CONVERA
CRIL TECHNOLOGY
DELPHES TECHNOLOGIES
DEMONIAK
DIAGONAL
DIALOCA
ELAN Informatique
EURASAM
EURODOC SOFILOG
EVER TEAM
GBConcept
GLOBALINK
HOLISTIQUE COMMUNICATIONS
IBM France
INFOGISTICS
iPHRASE
KOLTECH
LINGUATEC
LINGWAY
MACHINA SAPIENS
MEDIACONCEPT
MEMODATA
MIMETICS
MINDMAKER Inc
MONDOSOFT
MYSOFT
Multilingual Computing, Inc.
NAMING COMPANY LTD
NEMESIA
NEURO-CONCEPT
NEUROSOFT
NOEMATICS
NOMEN
Nomino Technologies Inc
PERTINENCE MINING Sarl
PROLOGIA
PROTEOR
RESUMIX
SAIL LABS
SDL International
SEMANTIA
SINEQUA
SOFTISSIMO
SYNAPSE
SYSTRAN
T-GID
TAGMATICA
TELISMA
TEMIS
TERAGRAM
TRIPLEHOP
VECSYS
VERITY FRANCE

[7] Cette liste ne saurait être exhaustive. Vous trouverez ses mises à jour sur le site de l'APIL.

Annexe B – Liste des références par type d'application[8]

Veille

Sociétés références	Fournisseur TAL	Urls Référence
CNES, EADS, FT , Alcatel, Pernod-Ricard, Total, St Gobain, DGA, Unilog, Fluxys	Arisem	www.arisem.com/fr/clients/index.html
Total	TEMIS	www.temis-group.com/temis/attachments/businesscases/Total_fr.pdf
Novartis	TEMIS	www.temis-group.com (lien *clients*)
Telecom Italia Mobile	TEMIS	www.temis-group.com (lien *clients*)

Bases de connaissance sémantiques

Sociétés références	Fournisseur TAL	Urls Référence
PSA, Hachette Filipacchi, Wolters Kluwer, Lexis Nexis	Mondeca	http://www.mondeca.com/fr/clients.htm

Portail

Sociétés références	Fournisseur TAL	Urls Référence
APEC	Lingway	www.lingway.com/commonfiles/Fiche_projet_APEC.pdf
INTEREX	Lingway	www.lingway.com/commonfiles/Interex_Fiche_projet.pdf
Total	TEMIS	www.temis-group.com/temis/attachments/businesscases/Total_fr.pdf
IPSEN	TEMIS	www.temis-group.com (lien *clients*)

[8] Ces références sont transmises par les entreprises qui ont fourni la solution TAL

Classification automatique

Sociétés références	Fournisseur TAL	Urls Référence
PSA	TEMIS	http://www.temis-group.com/temis/attachments/businesscases/PSA_fr.pdf
NOVARTIS	TEMIS	www.temis-group.com (lien *clients*)
TelCal (Telecom Calabria)	TEMIS	www.temis-group.com (lien *clients*)

Gestion des brevets

Sociétés références	Fournisseur TAL	Url Référence
QUESTEL ORBIT	Lingway	http://www.lingway.com/commonfiles/lingway_propriete_intellectuelle.pdf
NOVARTIS	TEMIS	www.temis-group.com (lien *clients*)
EADS CCR	TEMIS	www.temis-group.com (lien *clients*)
LION bioscience	TEMIS	www.temis-group.com (lien *partenaires*)

Terminologie d Entreprise

Sociétés références	Fournisseur TAL	Url Référence
Daimler Chrysler	TEMIS	www.temis-group.com/temis/attachments/businesscases/Daimler_fr.pdf

Ressources humaines

Sociétés références	Fournisseur TAL	Url Référence
Crédit Lyonnais, Département Ressources Humaines - Service Recrutement	KOLTECH TEMIS	www.focusrh.com/cgi-bin/site/site-newsview.pl?Cty=fr&Mail=031003icdbe3&News=04060358rxfo
Conoco (USA)	TEMIS	www.temis-group.com (lien *clients*)

Moteurs de recherche

Sociétés références	Fournisseur TAL	Urls Référence
C discount	SINEQUA	www.sinequa.com/html/temoignage-cdiscount.html
Leroy Merlin	SINEQUA	www.sinequa.com/html/temoignage-leroy-merlin.html
RSF (Reporters Sans Frontières)	SINEQUA	www.sinequa.com/html/temoignage-rsf.html
AMF (Autorité des Marchés Financiers)	SINEQUA	www.sinequa.com/html/temoignage-amf.html
Sénat	SINEQUA	www.sinequa.com/html/communique-34.html

Glossaire

APIL	Association des Professionnels des Industries de la langue
ASP	Application Service Provider ou Active Server Page selon le contexte
CRM	Customer Relationship Management
DSI	Direction des Systèmes d Information
ERMS	Electronic Response Management System
FTP	File Transfer Protocol
GED	Gestion Electronique de Documents
GFII	Groupement Français de l'Industrie de l'Information
GPS	Global Positioning System
IHM	Interface Homme Machine
ISO	International Standard Office
IVR	Integrated Voice Response (Reconnaissance indépendante du locuteur)
JVM	Java Virtual Machine
LMF	Lexical Markup Framework
MAF	Morpho-syntactic Annotation Framework
NTIC	Nouvelles Technologies de l Information et de la Communication
OCR	Optical Character Recognition (Reconnaissance Optique de caractères)
OWL	Ontology Web Language
RDF	Resource Description Framework
ROI	Return on Investment (Retour sur Investissement)
SQL	Structured Query Language
TAL	Traitement Automatique des Langues
TMF	Terminological Markup Framework
XML	eXtended Meta Markup Language

L'APIL tient également à remercier les entreprises mécènes pour leur soutien actif au domaine.

arisem

>lingway

MONDECA

SINEQUA

EMIS

BALISTE

Tagmatica

ILEDAC

Jean-François Delannoy

SOLACI

Index