île Rousse 2005
Journée sur les systèmes d'information élaborée

DEPLOIEMENT DES CONNAISSANCES LINGUISTIQUES DANS LES OUTILS DE VEILLE ET DE RECHERCHE SUR INTERNET : DES APPROCHES DIFFERENTES

Peggy Cadel,
ATER
IUT de Toulon - Département Services et Réseaux de Communication de Saint-Raphaël
Laboratoire I3M
cadel@univ-tln.fr
+33 4 94 19 66 00

Mots clés : Veille, recherche d'information, lemmatisation, troncature, traduction

Résumé
Parmi les connaissances linguistiques utilisées dans le cadre du traitement automatique des langues et plus précisément dans celui de la recherche d'information, d'un point de vue théorique, on distingue connaissances morphologiques, connaissances syntaxiques et connaissances sémantiques. D'un point de vue pratique, chacune de ces connaissances enrichissement les systèmes d'indexation et de recherche à différents niveaux. Tandis que les outils de veille et les outils de gestion de contenu s'arment de ces attributs afin de répondre à une réelle demande de la part des cellules de veilles et des unités de documentation, les outils de recherche sur Internet prennent une direction différente en présentant lemmatisation, troncature voire gestion des accents comme des arguments de recherche non pertinents et susceptibles de générer du bruit.

1 Introduction

L'étude des appels d'offre destinés à la mise en place de système de gestion d'information allant de l'acquisition à la classification et à la distribution ciblée de l'information, permet parallèlement aux contraintes techniques, architecturales, fonctionnelles et juridiques de mettre en lumière un besoin fort de traitements linguistiques. Connaissances morphologiques, syntaxiques et sémantiques deviennent ainsi incontournables pour les outils de veille sur Internet et les outils de gestion de contenu. De leur coté, les outils de recherche sur Internet semblent ne pas les intégrer lorsqu'il s'agit de la gestion de leur index mais s'y intéressent lorsqu'ils proposent des « outils de traduction ».
La description des connaissances linguistiques utilisées par les outils de veille et de gestion de contenu ainsi que le positionnement des moteurs de recherche seront présentés successivement.
Nous commencerons par décrire d'un point de vue fonctionnel les critères de recherche reposant sur des connaissances linguistiques essentiels pour les outils de veille et de gestion de contenu. Puis, nous nous intéresserons à la présence de ces critères chez les moteurs de recherche ce qui nous conduira à analyser leur positionnement linguistique.

2 Connaissances linguistiques nécessaires pour les outils de veille et de gestion de contenu

L'engouement pour les outils de veille et de gestion de contenu dits « intelligents » (ce qui se traduit d'un point de vue technologique par « manipulant des connaissances linguistiques ») par opposition aux outils fonctionnant à partir d'une méthode d'indexation uniquement « plein texte » (se limitant à la recherche des formes telles que saisies par l'utilisateur) résulte d'une double démarche.
La démarche initiale se concrétise par la volonté d'élargir les requêtes à des formes distinctes d'un point de vue graphique mais équivalentes d'un point de vue sémantique. Cette approche est liée à un constat d'impuissance de la part des utilisateurs conscients de la diversité d'expression d'un concept. Ils savent qu'ils peuvent se trouver face à des documents contenant les éléments informatifs qui les intéressent mais qu'ils n'y accèderont pas parce que l'information recherchée n'est pas exprimée par les mêmes termes à la requête et dans le texte. La volonté d'élargir les requêtes à des formes distinctes d'un point de vue graphique mais équivalentes d'un point de vue sémantique résulte de ce constat.
La seconde démarche a été initiée par les outils de veille et de gestion de contenu qui ont développé d'un point de vue technologique leurs modules linguistiques et d'un point de vue marketing leur argumentation « linguistique » en insistant sur les fonctionnalités de recherche pour soutenir les processus de traitement.
La couverture linguistique, transparente et fonctionnelle, est devenue nécessaire. Elle prend forme à travers les fonctionnalités de recherche suivantes :

- le traitement des formes fléchies (singulier / pluriel)
- la gestion de la troncature (la notion de linguistique n'est pas présente dans le processus mis en place mais dans l'utilisation du processus)
- l'élargissement des requêtes à partir de connaissances sémantiques
- les requêtes multilingues

.

2.1 Le traitement des formes fléchies

La prise en compte des formes fléchies est une contrainte fondamentale à laquelle aucun système ne peut plus se soustraire. Ils répondent d'ailleurs tous à ce besoin élémentaire soit, de manière élégante avec un lemmatiseur (impliquant l'utilisation d'un dictionnaire morphologique et d'un étiquetage syntaxique) soit, de manière plus brutale en utilisant un stemmeur (procédé nécessitant uniquement

l'emploi d'un fichier de formes fléchies fonctionnant à partir de l'extraction de terminaisons et de la concaténation de suffixes permettant la génération de nouvelles formes).
Le traitement des formes fléchies a l'avantage dans sa version la plus rudimentaire comme dans sa version la plus sophistiquée, d'être portable à tous types de recherche ou de manipulation d'information dans une langue donnée (ne dépend pas du domaine traité). Il offre à l'utilisateur la possibilité de trouver une forme différente au niveau morphologique mais identique au niveau sémantique de celle qu'il recherche.

2.2 La gestion de la troncature

La gestion de la troncature s'inscrit elle aussi au niveau morphologique. Ce procédé lui aussi assez rustre, permet à l'utilisateur de s'approprier sa requête en choisissant de l'élargir à n'importe quel suffixe, laissant la possibilité d'un rapprochement sous forme de famille lexicale en choisissant de ne pas figer la terminaison de la forme à rechercher. Cette extension de la requête non contrainte par un dictionnaire est aléatoire mais permet de s'abstraire de toutes contraintes formelles pouvant limiter la recherche.
La troncature est également utilisée dans le cas d'absence de traitement de formes fléchies et à l'intérieur des mots pour jouer le rôle d'un dictionnaire phonétique et suggérer différentes orthographes pour une forme (ex : Al ?da pour Al-Qaeda, AL-kaida…). Ce critère de recherche assez représentatif des pratiques de gestion électronique de documents est de plus en plus désiré par les utilisateurs des outils de gestion de contenu.

2.3 L'intervention de connaissances sémantiques

L'utilisation de connaissances sémantiques va offrir la possibilité d'élargir la recherche à des formes différentes de celles saisies. La finalité de cette opération est la mise en correspondance entre unités linguistiques représentant le même objet ou la même action (gestion des synonymes, des acronymes, des équivalences lexicales entre langues). Nous utilisons pour ce faire un thesaurus. Dans sa forme la plus simple le thesaurus se limite à un regroupement de synonymes hiérarchisés. Dans sa forme la plus complexe, le thesaurus permet de représenter l'ensemble des concepts du domaine et de leurs relations implicites.
Ce besoin trouve sa légitimité dans trois cadres spécifiques :

- une recherche d'information dans des domaines de spécialité pour lesquels une classification ne peut être remise en cause comme c'est le cas notamment dans le domaine médical avec le Mesh
- une application de veille, le concept de [concurrent] va être alors instancié de diverses manières en fonction des activités de l'acteur.
- Un accès à des sources d'informations multilingues.

2.4 Les requêtes multilingues

La possibilité de gérer des documents dans plusieurs langues n'est pas une fonctionnalité de recherche essentielle pour les utilisateurs des outils de veille et de gestion de contenu. Le souhait d'une gestion en parallèle de l'anglais et du français est le seul à apparaître mais ne représente pas une contrainte indispensable dans le choix de l'outil. En effet, si le besoin de surveiller et de gérer des sites dans plusieurs langues est réel, rares sont les prestataires d'outils de veille et de gestion de contenu qui ont choisi cette fonctionnalité de recherche comme un argument permettant de se distinguer de la concurrence. La technologie n'est donc pas encore transparente pour les utilisateurs qui ont comme image des systèmes multilingues les outils de traductions et leurs limites.

3 Connaissances linguistiques minimales pour les outils de recherche sur Internet

Les 3 principaux moteurs de recherche utilisés sont Google, Yahoo Search Technologie et Altavista. Nous ne nous intéresserons pas ici à Altavista que nous remplaçons par MSN Search. Le choix d'étudier MSN Search plutôt que Altavista se justifie pour deux raisons.
La première est la forte proximité commerciale et technologique existant entre Yahoo Search Technologie et Altavista.
La seconde est la récente mise en ligne du moteur de recherche de Microsoft et son fort potentiel à se positionner parmi les outils de recherche les plus utilisés si l'on se base sur le comportement des utilisateurs qui ont tendance à utiliser l'outil proposé en page d'accueil lors de leur connexion à internet. La page d'accueil proposée par défaut par le navigateur Internet Explorer est celle du site Web MSN sur lequel l'outil de recherche MSN Search est accessible.

L'étude des fonctionnalités de recherche des trois principaux moteurs de recherche sur Internet montre que leur approche des connaissances linguistiques est beaucoup plus complexe que les outils d'indexation et de recherche. Nous commencerons par exposer les fonctionnalités de recherche décrites dans la première partie puis nous nous interrogerons sur ces choix.

3.1 Fonctionnalités de recherche proposées

Le tableau suivant récapitule les différentes fonctionnalités de recherche proposées par Google, Yahoo Search Technologie et MSN Search :

Fonctionnalités de recherche proposées	**Google**	**Yahoo Search Technologie**	**MSN Search**
Traitement des formes fléchies	Non	Non	Non
Gestion de la troncature	Non	Non	Non
Utilisation de dictionnaires de synonymes ou de thesaurus	Oui[1]	Non	Non
Distinction entre majuscule et minuscule	Non	Non	Non
Distinction entre lettre accentuée et lettre non accentuée	Non	Oui et Non[2]	Non
Recherche dans plusieurs langues	Oui	Oui	Oui
Recherche dans plusieurs « pays »	Oui	Oui	Oui
Traduction des pages de résultats	Oui	Oui	Non

3.2 Choix des fonctionnalités de recherche

A première vue, aucune connaissance linguistique n'est utilisée et les formes recherchées sont les formes saisies. La « non extension » de la requête est-elle liée à la taille de la base interrogée (nombre de résultats, temps de réponse) ?
Aucun de ces acteurs n'évoquent de contraintes techniques mais Google justifie son choix par une volonté de précision :

[1] L'opérateur « ~ »permet de lancer la recherche sur des synonymes
[2] Le mot est toujours recherché tel qu'il est saisi sauf s'il n'est pas accentué.

> « Pour garantir des résultats aussi précis que possible, Google n'applique pas de « lemmatisation » (réduction des mots au masculin et/ou au singulier, à l'infinitif, etc.) et ne supporte pas les recherches à base de caractères joker/wildcard. Autrement dit, Google utilise les mots exactement tels que vous les entrez dans le champ de recherche. Exemple : Si vous entrez le terme « cheval » ou « cheva* », Google ne fait pas porter la recherche sur « chevaux », « chevaline » ou « chevalerie », mais uniquement sur le terme « cheval » ou « cheva* » (soit la chaîne de caractères « cheva » suivie d'un astérisque). Pour plus de sûreté, entrez les formes qui vous intéressent, par exemple : « cheval » et « chevalerie » .» http://www.google.fr/intl/fr/help/basics.html (citation extraite le 28/02/05)

L'emploi du terme « lemmatisation » dans l'aide de Google est assez inattendu, car ce document est un document non spécifique alors que le terme implique une connaissance particulière manipulée dans le cadre du traitement automatique des langues. Dans un même paragraphe, on constate également un « non emploi » du terme « troncature » qui consacre l'utilisation de « caractère joker ». Doit-on considérer que cette semi vulgarisation de processus linguistiques n'est présente que pour asseoir un discours rigoureux ?

Si l'approche est effectivement scientifique et la recherche de la forme exacte le fondement du module de recherche de Google alors comment expliquer la non distinction entre forme accentuée et non accentuée et entre majuscule et minuscule ? Ces distinctions ne permettent-elles pas aussi de préciser une requête ?
Cela signifie-t-il que l'utilisateur va délibérément choisir de poser une question au singulier ou au pluriel et donner du sens à ce choix mais qu'il ne donnera pas de portée sémantique au choix d'utiliser un caractère accentué ou non accentué et d'une majuscule ou d'une minuscule ? Les orientations linguistiques de tel ou tel moteur sont-elles en réalité plus portées par des contraintes techniques que par de réelles théories linguistiques ? Est-ce pour cette raison qu'à l'exception de Google, les autres moteurs de recherche restent laconiques sur leurs choix ? Doit-on voir dans cette volonté de justification linguistique une tentative de rapprochement avec certains traitements linguistiques comme l'élimination de mots vides présentée ci-dessous ?

> « Google ignore les chaînes de caractères dont le poids sémantique est trop faible (également désignés « mots vides » ou « bruit ») : le, la, les, du, avec, vous, etc., mais aussi des mots spécialisés tels que « http » et « .com » et les lettres/chiffres d'un seul caractère, qui jouent rarement un rôle intéressant dans les recherches et risquent de ralentir notablement le processus.
> Pour forcer l'inclusion d'un mot vide dans une recherche, il suffit de le faire précéder du signe plus (+), lui-même précédé d'un espace. [Vous pouvez également inclure le signe plus (+) dans une recherche d'expression, par exemple « +sur +le perron +de +la femme +du boulanger +de Nevers » .] » http://www.google.fr/intl/fr/help/basics.html (citation extraite le 28/02/05)

Là encore, le discours de Google se veut « scientifique », il est question de chaîne de caractère, de « poids sémantique » et de « mots vides ». Le discours de MSN Search à ce sujet reprend un vocabulaire moins technique en évoquant les « mots courants » mais beaucoup plus pratique en évoquant la réalité d'interrogation liée à l'usage des « mots clés » OR et NOT :

> « Les mots courants comme un, et, ou et le sont ignorés sauf si vous les mettez entre guillemets. (Les mots clés OR et NOT servent à combiner ou à exclure des termes.) » http://search.msn.fr/docs/help.aspx?t=SEARCH_GS_ABCsOfFindingInformation.htm
> (citation extraite le 05/05/05)

Afin de compléter notre observation lexicale des explications fournies par les moteurs de recherche, il est intéressant de noter que le commentaire de MSN Search est introduit par le titre « Utilisateurs novices : recherche d'informations sur le Web » ce qui démontre une réelle volonté de s'adapter à un public non spécialiste. Cette adaptation prend tout son sens sur la page intitulée « Quatre principaux problèmes rencontrés au cours de recherches sur le Web » lorsqu'une acquisition d'informations insuffisantes est ainsi décrite :

> « Lorsque vous menez votre voiture à réparer, vous ne pouvez pas vous contenter de dire au mécanicien « Ma voiture a des problèmes » ; vous devez être plus spécifique. Il en va de même avec les moteurs de recherche. Si vous tapez problèmes de voiture dans la zone de recherche, vous allez obtenir un grand nombre de résultats. Mais, si vous spécifiez la marque, le modèle et l'année de votre véhicule, accompagnés d'une description du problème, vous aurez plus de chances de trouver les informations dont vous avez besoin. »
> http://search.msn.fr/docs/help.aspx?t=SEARCH_TROU_Top4ProblemsSearchingWeb.htm
> (citation extraite le 10/05/05)

La « vulgarisation » de la recherche d'information proposée par MSN Search trouve ses limites dans son ergonomie, l'aide proposée est composée de nombreuses pages Web très denses dont le contenu dépasse souvent la « page écran ».
Entre l'utilisation de termes techniques de Google présentée de manière conviviale et les explications vulgarisées de MSN Search noyées dans une page surchargée, Yahoo Search Technologie se positionne en intermédiaire reprenant la présentation de Google et la simplicité de MSN Search sans cependant fournir de réels éléments informatifs sur les processus de traitement qu'il utilise.

3.3 Les requêtes multilingues

Le rapprochement entre moteur de recherche sur Internet et connaissances linguistiques semble associé au choix de nommer les fonctionnalités de traduction : « outils linguistiques » ?
Outils linguistiques et choix des langues ne signifient pas forcément utilisation de connaissances linguistiques. Lorsqu'il est possible d'effectuer une recherche sur un pays, la recherche est effectuée uniquement sur un nom de domaine (ex : .fr, .it…). Lorsque qu'il s'agit de la langue dans laquelle le texte est écrit (si l'on reprend les termes de yahoo), nous sommes effectivement dans le cadre de la reconnaissance linguistique gérée par un module se basant sur les chaînes de caractères les plus utilisées par chaque langue.

Une fois encore si l'on s'intéresse aux processus de traduction, l'acteur le plus locace sur les processus utilisés est Google.

> « La traduction proposée par Google a été générée automatiquement par la technologie la plus perfectionnée qui soit disponible actuellement. Malheureusement, force est de reconnaître que les logiciels les plus évolués du marché sont bien loin d'égaler la maîtrise linguistique d'une locutrice native ou les compétences d'un traducteur professionnel. La traduction automatique reste une technologie très complexe et difficile à implémenter, dans la mesure où le sens d'un mot peut varier en fonction du contexte dans lequel il est utilisé (littéraire ? scientifique ? religieux ?), du niveau de langage pratiqué (spécialiste ? débutant ?), des interactions rédacteur-lecteur (style pompeux ? accessible ?) et de nombreuses références socioculturelles. Pour toutes ces raisons, une traduction précise (ou au moins correcte et compréhensible) exige que l'entité intervenante -- personne ou logiciel – maîtrise le contexte, la structure et les règles de la paire de langues considérée. Les ingénieurs et les linguistes se penchent sur ce problème depuis plusieurs dizaines d'années, mais il faudra sans doute plusieurs lustres (et quelques lumières...) avant qu'il soit possible de disposer d'un système automatique de traduction dont les résultats sont comparables à ce que produit un cerveau humain spécialisé dans le domaine. Dans l'attente, nous espérons que le service de traduction que nous vous proposons est satisfaisant dans la plupart des cas. N'hésitez pas à nous donner votre opinion. »
> http://www.google.com/intl/fr/help/faq_translation.html
> (citation extraite le 09/03/05)

Cette présentation des contraintes liées à la traduction automatique est remarquable d'honnêteté mais ne nous permet pas d'émettre une quelconque hypothèse quant aux formalismes de traduction utilisés. Le seul élément informatif qui apparaît dans les propositions de traduction est le choix du couple

langue source / langue cible. Le choix proposé présente des couples de langue avec comme langue cible ou source obligatoire l'anglais ce qui peut signifier que l'anglais joue le rôle de langage pivot permettant de représenter à la fois les contraintes intralinguistiques propres à cette langue mais aussi les contraintes extralinguistiques communes à l'ensemble des langues. Il est à noter que le système ne propose pas dans sa version actuelle de traduction de langues à alphabets cyrilliques mais qu'il traite des langues asiatiques ce qui laisse supposer que d'un point de vue technique le traitement des langues à alphabet cyrillique ne pose aucune difficulté.

Même si l'outil linguistique de traduction choisi par Yahoo Search Technologie, n'est pas aussi documenté que celui de Google, il nous est possible d'expliquer les fonctionnements de ces procédés de traduction car il utilise une technologie développée par un des pionniers en matière de traduction : Systran. Systran procède par couple de langues sans s'aider d'une représentation extralinguistique. Cette méthodologie à l'origine des premiers systèmes de traduction s'est accompagnée de célèbres écueils aujourd'hui compensés par une analyse plus fine du co-texte.

Alors que Yahoo Search Techonologie et Google se sont dotés d'une fonctionnalité de traduction, MSN Search se contente de proposer des recherches par langue et par pays. Cette attitude est paradoxale au vu de l'expérience dont dispose Microsoft en matière de traitement automatique des langues : correcteur orthographique, correcteur grammatical, dictionnaire de synonymes.

4 Perspectives

Les outils de recherche sur Internet ne se préoccupent pas des traitements linguistiques élémentaires mais s'intéressent à des processus aussi complexes que la traduction ou le rapprochement avec des encyclopédies.
Par conséquent, nous pouvons nous demander si la non gestion des formes fléchies des outils de recherche sur Internet est liée à des questions de taille de la base de connaissance, de sa structuration ou d'une volonté de ne pas accroître la masse d'information disponible pour l'internaute ? La taille du corpus doit-elle alors être considérée comme un critère déterminant dans le choix du système d'indexation et l'approche plein texte comme la méthode d'indexation pour les gros corpus,
Les trois moteurs de recherche semblent adopter les mêmes choix technologiques et s'intéresser au même titre à l'encyclopédie en ligne Wikipedia, de type "open source", libre d'accès et d'utilisation.
Google a proposé des serveurs et de la bande passante à Wikipedia et Yahoo fournit également du matériel et des ressources informatiques à Wikipedia, ces rapprochements qu'en partie altruistes s'accompagnent d'un accès à l'encyclopédie pour les utilisateurs des deux moteurs. MSN Search quant à lui propose un accès gratuit à l'encyclopédie payante Encarta mais figure aux côtes de Google et de Yahoo Search Technologie sur le site de Wikipedia lors de l'élargissement de la requête à des documents externes.
Les outils de recherche sur Internet ont un positionnement ambigu face aux connaissances linguistiques alors que les outils de veille et de gestion de contenu ont grâce un discours non équivoque réussi à les imposer. Nous pouvons cependant nous demander si les liens qui sont en train de se tisser entre le domaine de la linguistique et de la recherche d'information ne vont pas remettre en question cette dichotomie.

5 Bibliographie

BACHIMONT B., 2003, L'indexation multimédia in GAUSSIER E. et STEFANINI M-H, Assistance intelligente à la recherche d'informations (Traité STI), p153-184

BRANCIER C., 2004, Décision Micro, Veille sur le Net : des besoins variés
Http://www.01net.com/article/240376.html

BRUANDET M-F, CHEVALET J-P, 2003, Utilisation et construction de bases de connaissances pour la recherche d'information in GAUSSIER E. et STEFANINI M-H, Assistance intelligente à la recherche d'informations (Traité STI), p99-132.

DOU H., 1995, Veille technologique et compétitivité, Dunod, Paris.

MEINGAN D., LEBO I., 2004, Livre blanc : "Maîtriser la veille pour préparer l'intelligence économique"
http://www.knowledgeconsult.com/fr/article.php3?id_article=37

QUESTER Christophe, 2004, Solutions : les Français maîtrisent le terrain, 01 DSI
http://www.01net.com/article/262627.html

SAMIER H., SANDOVAL V.,1999, La recherche intelligente sur l'Internet et l'intranet, 2ème édition revue et augmentée, Hermès, Paris.