L'apport de connaissances linguistiques en recherche documentaire

Claude de Loupy

(1) Sinequa[1]
51, rue Ledru-Rollin Ivry-sur-Seine
loupy@sinequa.com

Résumé – Abstract

L'utilisation de connaissances et de traitements linguistiques évolués en recherche documentaire ne fait pas l'unanimité dans le milieu scientifique. En effet, de nombreuses expériences semblent montrer que les résultats obtenus ne sont pas améliorés, voire sont parfois dégradés, lorsque de telles connaissances sont utilisées dans un système de RD. Dans ce tutoriel, nous montrons que les environnements d'évaluation ne sont pas adaptés aux besoins réels d'un utilisateur car celui-ci recherche presque toujours une information. Il veut donc retrouver des documents pertinents le plus rapidement possible car ce n'est pas là le but de sa recherche. Le temps global de la recherche est donc fondamentalement important. Néanmoins, le cadre d'évaluation TREC nous permet de montrer que l'utilisation de connaissances linguistiques permet d'augmenter la précision des premiers documents renvoyés, ce qui est très important pour diminuer le temps de recherche.

The use of linguistic knowledge and natural language processing (NLP) systems is generally not considered to be the best way to perform document retrieval. Indeed numerous experiments seem to show that NLP does not improve the performances and, even can decrease them. In this tutorial we show the environments used for evaluation do not fit the real world needs. Almost every time, the user search for an information. He wants relevant documents as soon as possible because they are not his goal. So, time is really important in evaluation of document retrieval systems. Nevertheless, using TREC, we show NLP can lead to a better precision for the first documents retrieved and this is really important in order to improve the speed of search process.

---

[1] Beaucoup des résultats présentés ici ont été obtenus alors que l'auteur était au Laboratoire Informatique d'Avignon ou à la fois au Laboratoire Informatique d'Avignon et chez Bertin Technologies.

1 Introduction

Parmi les principales difficultés rencontrées par les moteurs de recherche, celles posées par les langues elle-mêmes sont probablement les plus difficiles. C'est pourquoi, de nombreuses expériences utilisant des ressources ou traitements linguistiques ont été effectuées et présentées dans la littérature. Pourtant, même si certains résultats sont encourageants, l'utilité de tels traitements n'a pas encore été montrée de manière claire. Le but de ce tutoriel est de présenter un certain nombre de ces difficultés et les méthodes utilisables pour les traiter. Nous montrons aussi que si les résultats rapportés sont souvent négatifs, c'est parce que les environnements d'évaluations ne font pas ressortir les points les plus importants.

Nous présentons tout d'abord le cadre d'évaluation TREC avec lequel certains résultats ont été obtenus. Ensuite, nous évoquons certaines difficultés liées aux langues ainsi que les possibilités de traitement qui permettent de les surmonter en les illustrant par des expériences que nous avons menées sur TREC. Nous montrons aussi que l'utilisation de connaissances linguistiques est nécessaire dans le cadre d'une interaction avec l'utilisateur. Enfin, nous montrons que le point de vue classique, illustré par TREC ne suffit pas à évaluer un système de recherche documentaire et que la considération des points les plus importants montre l'intérêt de la linguistique pour augmenter la vitesse de la recherche.

2 L'évaluation des systèmes

2.1 Les environnements d'évaluation

Afin de répondre au mieux aux attentes de l'utilisateur, il est nécessaire de pouvoir évaluer les performances des systèmes de recherche documentaire. Mais il est très difficile d'effectuer ces évaluations car la satisfaction de l'utilisateur, pourtant fondamentale, ne peut pas être évaluée de manière certaine. Plusieurs méthodes permettant de mesurer l'efficacité d'un système de recherche documentaire existent. Mais la mise en place d'environnements d'évaluation et même l'interprétation des résultats obtenus ne sont pas triviales (Loupy & Bellot, 2000).

Afin de pouvoir évaluer un système de recherche documentaire, il convient déjà de disposer d'un cadre d'évaluation où les requêtes sont données et les documents pertinents qui leur correspondent connus. Il existe de nombreux environnements d'évaluation des systèmes de recherche documentaire. Le plus important d'entre eux est TREC (Text REtrieval Conference) (Harman, 1993). Cette campagne de grande envergure a lieu chaque année. Chaque évaluation comporte 50 requêtes qui doivent être appliquées à un corpus dont la taille est d'environ 1,9 Go pour 528 155 documents.

Une requête TREC, telle qu'elle est donnée aux participants, comporte différents éléments de taille croissante : un titre (un petit nombre de mots indiquant le thème recherché), une description (une ou deux phrases permettant de mieux cerner le sujet) et une explication (ici, la requête est expliquée en détail avec un grand nombre de termes liés, mais aussi de négations du type « *les documents rapportés ne devront pas parler de...* »). Il nous paraît plus intéressant de centrer les expériences et les résultats sur des requêtes courtes. En effet, Croft *et al.* (1995) rapportent que la longueur moyenne des requêtes posées à la base *Thomas*

(textes législatifs) est de 2,31 mots. Nous avons donc utilisé ce qui est donné dans la rubrique *Title* des requêtes TREC. Ces requêtes ont entre 1 et 4 mots. Cela rejoint le sondage effectué par abondance (abondance, 2000) avec la question : « Lorsque vous utilisez un outil de recherche, combien de mots clés tapez-vous en moyenne ? »

1 mot clé	11,47%
2 mots clés	57,80%
3 mots clés	24,31%
4 mots clés et plus	0%

Figure 1 : Sondage Abondance sur le nombre de mots par requête.

2.2 Rappel et précision

Deux indicateurs qui permettent de rendre compte de la qualité globale de la réponse d'un système à une requête sont couramment utilisés :

- La ***précision*** : la précision correspond au pourcentage de documents pertinents renvoyés par le système qui répondent effectivement à la requête.

$$précision = \frac{nb.\ documents\ pertinents\ renvoyés}{nb.\ documents\ renvoyés\ par\ le\ système}$$

- Le ***rappel*** : le rappel désigne le pourcentage de documents pertinents rapportés par le système par rapport au nombre total de documents pertinents qui se trouvent dans la base documentaire.

$$Rappel = \frac{nb.\ documents\ pertinents\ renvoyés}{nb.\ documents\ pertinents\ dans\ le\ corpus}$$

2.3 L'évaluation « à la TREC »

L'évaluation TREC se fait en fonction d'un certain nombre de critères :

- Le nombre de documents pertinents trouvés par le système.
- Les résultats, en terme de précision, correspondant à des valeurs fixes de rappel (précision pour un rappel de 0,1, pour un rappel de 0,2, etc.).
- La précision moyenne qui correspond à la somme de la précision pour chaque document pertinent rapporté divisée par le nombre total de documents pertinents.
- La précision pour *n* documents rapportés (précision pour 5, 10, 15, 20, 30, 100, 200, 500 et 1 000 documents rapportés).
- La précision pour un nombre de documents rapportés égal au nombre de documents pertinents présents dans la base, notée *R-Prec*.

3 Quels problèmes posent les langues aux moteurs de recherche ?

Il existe bien des difficultés à surmonter pour créer un système de RD efficace : rapidité de l'indexation et de la recherche, taille de l'index, robustesse, fiabilité, efficacité, etc. Mais les problèmes les plus difficiles ne correspondent pas à de la technique pure. Ils sont liés aux propriétés même des langues. Nous évoquons certaines des difficultés les plus importants.

3.1 Les niveaux graphique, morphologique, grammatical et syntaxique

3.1.1 La graphie

Un mot peut s'écrire de plusieurs façons (Khadaffi, Kaddafi, etc.), comporter des fautes de frappe ou d'orthographe ou s'écrire avec une majuscule. Cela diminue le rappel car si un mot est orthographié d'une certaine façon dans la requête (*Khadaffi*), la simple recherche de ce mot ne permet pas de retrouver les documents qui le contiennent sous une autre forme graphique (*Kaddafi*). Dans le cas de variantes graphiques, il est possible d'utiliser un module de phonétisation. Cela permet de rapprocher des graphies en fonction de leur prononciation : d'après les règles de prononciation française, Khadaffi et Kaddafi sont équivalents. En ce qui concerne les fautes de frappe ou d'orthographe, il est possible d'utiliser des heuristiques propres aux correcteurs orthographiques (phénomènes d'insertion, de répétition, de dédoublement, etc.) qui peuvent d'ailleurs aussi faire intervenir de la phonétique.

D'un autre côté, la recherche systématique des variantes d'un mot diminue la précision ; ce n'est pas parce qu'un mot inconnu est proche d'un autre mot qu'il a été mal orthographié. De plus, dans certains cas, la graphie peut donner une information relative au sens du mot utilisé (Loupy & El-Bèze, 2000). C'est par exemple le cas pour le mot *Histoire* dont la majuscule indique qu'il s'agit de la discipline étudiant le passé et non d'une anecdote.

3.1.2 Les variantes grammaticales

Une même forme (*plane*) peut être verbe ou adjectif. Identifier le terme *plane* en tant qu'adjectif dans une requête permet d'écarter les textes dans lesquels *plane* apparaît en tant que verbe. Il y a donc augmentation de la précision. L'utilisation d'un outil d'étiquetage grammatical dans un système de traitement de RD devrait permettre d'obtenir un gain de performance en terme de précision (Hull *et al.*, 1996 ; Losee, 1996). En effet, il y a plusieurs avantages supplémentaires à connaître la catégorie grammaticale d'un mot en contexte :

- Le premier est de distinguer les mots outils des mots porteurs de sens. Dans une requête telle que « *La croissance économique* », l'article *la* a peu d'importance. Des anti-lexiques[2] ont donc été constitués afin d'éliminer ces mots trop fréquents lors de la phase d'indexation (Fox, 1990). Cela permet de ne pas avoir à traiter des mots qui n'apportent que peu d'information. Mais l'élimination systématique de l'article *la*

[2] *stoplist* en anglais.

conduit à éliminer un terme important en musique (la note *La*). Il est donc préférable de se baser sur un étiquetage grammatical plutôt que sur la simple graphie.

- En deuxième lieu, la connaissance d'un mot et de sa catégorie grammaticale permet, de manière quasi-systématique, de connaître son lemme, c'est à dire une forme normalisée. Ainsi, la connaissance de la qualité de substantif d'une occurrence du mot *tables*, permet de mettre en relation ce dernier avec le lemme *table*, plutôt qu'avec le lemme *tabler* (verbe). Ce rattachement du mot à sa racine permet de résoudre en grande partie le problème des variations morphologiques évoqué dans la section précédente.

- De plus, la connaissance du lemme et de la catégorie grammaticale est un élément essentiel pour la désambiguïsation sémantique (Wilks, Stevenson, 1996).

3.1.3 Les variations morphologiques

Les marques de nombre, de genre, les conjugaisons, etc. modifient la graphie d'un terme. Ce phénomène diminue le rappel des systèmes de RD. Il faut pouvoir retrouver la forme *chevaux* si la requête comporte le nom *cheval*. Pour cela, il est possible d'utiliser une procédure qui retrouve la racine linguistique (ou lemme) du terme appelée lemmatisation. Cette procédure se fait après un étiquetage grammatical.

On oppose souvent à la procédure de lemmatisation, la procédure de stemming qui consiste à rechercher, pour un terme donné, une pseudo-racine appelée *stem*. Le but de ce traitement est de ramener un terme à une de ses parties qui le caractérise ainsi que tous les termes qui lui sont linguistiquement liés. Ainsi, le stem de *déménageur* serait *déménag*. Ce dernier définit un groupe comprenant, entre autres {*déménageur*, *déménageurs*, *déménagement*, *déménagements*, *déménager*, *déménage*, etc.}. On voit donc bien ici l'apport d'une telle opération puisque des termes fortement liés grammaticalement et sémantiquement sont regroupés dans un même ensemble. Il ne s'agit plus seulement de la prise en compte de la morphologie flexionnelle mais aussi de la morphologie dérivationnelle. Une procédure de lemmatisation n'aurait pas permis de trouver un lien entre *déménagement* et *déménageur*. Malheureusement, la construction des stems n'est quasiment jamais basée sur des contraintes linguistiques fortes. Leur utilisation conduit donc souvent à une augmentation du bruit. Ainsi, le stem *port* renvoie aussi bien au mot *port* (pour les navires) qu'à toutes les formes du verbe *porter* alors qu'aucun lien linguistique ne lie ces deux termes. De plus, une opération simple de stemming ne pourra jamais faire correspondre *œil* avec son pluriel *yeux*.

Bien que le stemming permette souvent de retrouver des termes fortement liés sémantiquement ou grammaticalement, des confusions apparaissent souvent et apportent une information préjudiciable aux performances. Il convient donc de manipuler cet enrichissement avec prudence. Des expériences[3] conduites en utilisant TREC-6 permettent de comparer les performances d'un système utilisant une lemmatisation ou une opération de stemming. Le

---

[3] Le système de RD utilisé est le système IndeXal de Bertin Technologies (Loupy et al., 1998a). Pour l'étiquetage grammatical et la lemmatisation, nous avons utilisé le système ECSTA (Spriet & El-Bèze, 1997).

tableau suivant présente la précision obtenue pour différentes valeurs de rappel et selon le nombre de document rapportés ainsi que la précision moyenne et la R-précision.

	Orig.	**Stem.**	**Lem.**	**Lem.-Stem.**
Nb. doc. pertinents retrouvés	1841	2068	1987	2124
Rappel=0,1	41,2 %	44,5 %	45,0 %	46,4 %
Rappel=0,2	28,8 %	35,0 %	35,3 %	37,8 %
Rappel=0,5	18,0 %	21,6 %	19,8 %	22,7 %
Précision moyenne	18,7 %	22,3 %	21,1 %	23,1 %
Nb. docs=5	33,6 %	38,4 %	40,0 %	39,2 %
Nb. docs=10	32,0 %	34,8 %	36,0 %	36,0 %
Nb. docs=20	28,5 %	31,9 %	30,8 %	31,7 %
Nb. docs=100	15,4 %	17,2 %	16,8 %	17,9 %
R-Précicion	22,6 %	26,1 %	25,6 %	26,9 %

Figure 2 : Comparaison d'une lemmatisation et d'une opération de stemming

On peut constater que, globalement, les performances des lemmes sont moins bonne que celle des stems. La différence dans les performances a deux causes principales : l'absence de liens entre catégories grammaticales lors de l'utilisation des lemmes et la gestion des mots inconnus. En effet, dans le cas de la requête *325* (« *Cult lifestyles* »), le terme *lifestyles* est inconnu du lexique utilisé. Il a donc été impossible de retrouver des documents contenant *lifestyle* au singulier. Ainsi, la précision moyenne pour cette requête est de 8,4 pour les lemmes et de 16,2 pour les stems. Il est à noter que l'utilisation d'un module de traitement des mots inconnus aurait sans doute permis de rattacher *lifestyles* à son singulier *lifestyle*, ce qui permettrait d'éviter de recourir à une procédure de stemming dans ce cas (Spriet *et al.*, 1996).

Mais une information fondamentale est que les performances sont meilleures en tête de liste (jusqu'à 10 documents rapportés). Nous y reviendrons par la suite (cf. section 5). On peut aussi constater qu'il est possible de profiter des avantages des deux procédures en les combinant. Cette combinaison obtient quasi-systématiquement de meilleurs résultats que les deux procédures prises séparément.

Si l'on peut constater une amélioration nette des performances en utilisant un analyseur morphologique (même de bas niveau), il convient tout de même d'utiliser un tel module avec précautions car cela peut, dans certains cas, baisser la précision (Krovetz, 1993). Par exemple, certains traits morphologiques donnent une information sémantique (Loupy & El-Bèze, 2000). Ainsi, ramener la forme *eaux* à sa racine *eau* peut signifier une perte d'information sémantique si la requête concerne les « *eaux chez une femme enceinte* ».

3.1.4 La composition lexicale

La prise en compte des expressions composées devrait permettre l'augmentation de la précision : reconnaître l'expression *pied de biche* permettrait d'éviter de considérer pertinent un texte contenant « *la biche était couchée à ses pieds* ». En particulier, dans les domaines spécialisés, l'information pertinente est souvent contenue dans les groupes nominaux. De plus, le nombre de mots composés, pour le français, est largement supérieur à celui des mots simples (Royauté *et al.*, 1992).

La simple prise en compte de la distance entre les termes au sein d'un document permet d'augmenter la précision des réponses de manière importante : on passe de 39,6 % de bonnes réponses pour les 5 premiers documents à 44,4 % quand on prend en compte cette proximité (Loupy, 2000).

Certaines langues comme l'allemand sont agglutinantes. Il est alors nécessaire de segmenter les unités composées afin de retrouver les termes unitaires. Le procédé est alors inverse de celui de la recherche d'expressions dans une suite de mots séparés.

En revanche, une recherche portant sur les *capteurs de température* doit permettre de retrouver la phrase « *La température peut être mesurée à l'aide de capteurs* ». Il faut donc veiller à ne pas trop privilégier la recherche d'expressions en laissant toujours la possibilité de la recherche par mots simples.

3.1.5 La prise en compte de la syntaxe

La gestion d'informations syntaxique permet principalement de déterminer quels sont les groupes nominaux d'une phrase ainsi que les sujets et compléments des verbes présents dans les textes. Les expériences effectuées avec de telles analyses ne montrent pas de gain substantiel dans le cadre d'évaluation type TREC. Ce genre de traitement est surtout utile pour des besoins type extraction d'information, réponse précise à une question.

3.2 Le niveau sémantique

3.2.1 La synonymie

Lorsqu'une recherche est faite sur un mot comme *miroir*, il faut pouvoir retrouver des textes utilisant des synonymes de ce mot (*glace*), pour accroître le rappel. Ce phénomène est fortement lié à la polysémie, car il est indispensable de déterminer le sens dans lequel le mot est employé avant de rechercher ses synonymes. Par exemple, si on ne détermine pas le sens du terme *circulation* dans l'expression « *la circulation sanguine* », le risque est d'enrichir la requête avec des termes comme *trafic* et donc d'augmenter considérablement le bruit. De plus, lors de l'enrichissement, il convient de limiter la recherche du terme d'enrichissement à ses occurrences dans le sens recherché. Dans le cas de la relation de synonymie *miroir* / *glace*, il ne faut pas rapporter des documents qui parlent de « *glace à la vanille* » ! D'autres relations sémantiques entrent en jeu comme l'hyponymie (relation père / fils).

3.2.2 La polysémie

Si une requête contient l'expression « *Table de logarithmes* », le terme *table* ne doit pas être mis en relation avec des documents qui utilisent ce terme pour désigner une « *table de cuisine* ». Ce phénomène est très courant car beaucoup de mots sont polysémiques et il a pour effet de diminuer la précision des systèmes. Mais la levée de l'ambiguïté sémantique est très difficile ainsi que l'a montré la campagne d'évaluation Senseval (Kilgarriff, Palmer, 2000). Les meilleurs systèmes ont des performances inférieurs à 80 % de bon étiquetage. Bien sûr, il convient de relativiser cette évaluation car certaines distinctions de sens ne sont pas utile lors

d'une recherche documentaire. Ainsi, il paraît totalement inefficace de considérer les 41 sens que WordNet donne au verbe *run* dans le cadre d'une recherche documentaire (Gonzalo et al., 1998 ; Palmer, 1998).

Nous avons utilisé un système de désambiguïsation sémantique (Loupy et al., 1998b) afin de tester l'influence d'un tel module sur les performances de l'outil. Cet outil permet, dans un premier temps, d'affecter un sens aux mots des documents au moment de l'indexation et aux termes de la requête au moment de la recherche. Une fois le sens connu, il est alors possible d'utiliser un enrichissement par synonymie. Le tableau suivant présente les résultats obtenus par une lemmatisation simple, en n'enrichissant que les termes dont le sens n'est pas ambigu, en désambiguïsant le sens des termes et en couplant à la fois une désambiguïsation et un enrichissement à l'aide des synonymes une fois le sens connu.

	Lem.	**Syn.**	**désamb.**	**désamb.+syn.**
Nb. doc. pertinents retrouvés	1987	1999	1976	1973
Rappel=0,1	45,0 %	45,1 %	43,9 %	45,0 %
Rappel=0,2	35,3 %	35,5 %	35,0 %	35,1 %
Rappel=0,5	19,8 %	21,0 %	19,6 %	19,7 %
Précision moyenne	21,1 %	21,4 %	20,9 %	21,1 %
Nb. docs=5	40,0 %	39,6 %	41,2 %	42,8 %
Nb. docs=10	36,0 %	36,2 %	36,0 %	36,2 %
Nb. docs=20	30,8 %	30,9 %	30,3 %	30,1 %
Nb. docs=100	16,8 %	16,8 %	16,8 %	16,8 %
R-Précicion	25,6 %	25,7 %	25,1 %	25,4 %

Figure 3 : Évaluation des performances en utilisant un enrichissement sémantique et une désambiguïsation sémantique.

On peut constater plusieurs choses à partir de ce tableau :

- Contrairement à ce qu'affirme Sanderson (1994), il n'est pas nécessaire que les performances de la désambiguïsation sémantique soit supérieure à 90 % pour que les performances du système de RD ne soient pas baissées. Ici, nous avons une désambiguïsation dont nous avons évalué les performances à 71,5 % et dont l'application à la RD donne des résultats équivalents (même s'ils sont très légèrement inférieurs) à une simple lemmatisation.

- Il convient de remarquer aussi que le rappel n'est presque pas modifié lorsqu'on applique la désambiguïsation sémantique. En ne conservant que le premier sens, seules 13 requêtes voient le nombre de documents pertinents ramenés modifiés (8 à la baisse et 5 à la hausse). Il s'agit donc, principalement d'un ré-ordonnancement des documents. Cela permet de conclure d'une part que, contrairement à ce qui est communément admis, la désambiguïsation sémantique ne fait pas baisser le rappel et, d'autre part, que la désambiguïsation est implicitement faite par la présence des autres termes de la requête.

- Là aussi, on peut constater une amélioration non négligeable du pourcentage de textes pertinents dans les premiers renvoyés par le système lorsqu'on utilise à la fois la désambiguïsation et l'enrichissement par les synonymes.

3.2.3 La pragmatique

3.2.3.1 Connaissances spécialisées

Selon le niveau des traitements linguistiques effectués, il peut être beaucoup plus efficace d'utiliser des connaissances spécifiques au domaine ciblé, plutôt que des données généralistes. Cela suppose de disposer d'un lexique spécialisé pour le domaine traité. Par exemple, les termes utilisés en navigation n'ont pas forcément le même sens que dans la couture (*voile*).

Des expériences ont été faites sur les 10 premières requêtes de TREC-6 (Loupy, 2000). Un thésaurus spécialisé a été construit manuellement pour les thématiques concernées par ces requêtes (sans tenir compte des termes utilisés dans l'ensemble de la requête TREC). L'utilisation de ces thesaurus spécialisés permet d'obtenir un gain de 5 % (en absolu) pour la précision à 20 documents rapportés et de 10 % dans le cas de deux requêtes. Pour les 4 autres requêtes, les performances sont inchangées. En particulier, la requête 310 (« African civilian deaths ») voit sa précision moyenne augmenter de plus de 11 % en absolu grâce à l'utilisation d'un thesaurus géographique donnant les pays, noms d'habitants et grandes villes d'Afrique.

3.2.3.2 Découpage thématique

Sinequa utilise, dans son système *Intuition*, un lexique sémantique général dans lequel chaque terme est lié à un ou plusieurs domaines appelés descripteurs. Ces descripteurs correspondent à des « sacs de mots » (« bags of words ») pour lesquels un lien thématique a été établi. Ainsi, le descripteur concernant la « main » comprendra des termes comme *main*, *ongle*, *paume*, etc. Ainsi, l'outil met en correspondance non seulement les mots mais aussi les thèmes d'une requête et de documents. Lors de l'indexation, les mots utilisés dans un document sont mis en correspondance avec les descripteurs afin de déterminer les thématiques saillantes du texte. Les requêtes sont traitées de la même façon. En général, une seule thématique en ressort, mais il est possible à l'utilisateur de préciser le domaine. Par exemple s'il pose le seul terme *avocat* comme requête, il lui est possible de choisir entre les domaines *botanique* et *justice*.

Nous avons utilisé ce lexique dans le cadre d'un système de classification automatique de mails. Un environnement d'évaluation a spécialement été développé par le client : 20 000 mails associés à une des 120 catégories. La figure suivante présente la courbe des performances selon le degré d'utilisation de la sémantique. Pour un *alpha* nul, seuls les mots sont utilisés et pour un *alpha* égal à 100, seuls les descripteurs sont utilisés.

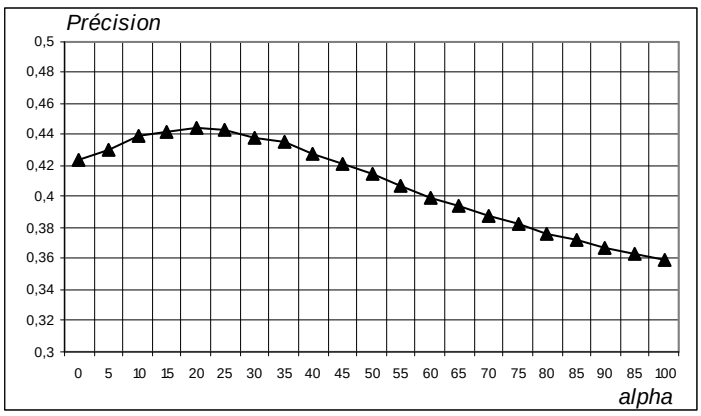

Figure 4 : Apport de connaissances thématiques pour le routage de mails.

Sur cette courbe, on voit que l'utilisation de la sémantique permet d'améliorer sensiblement les résultats par rapport à la seule utilisation des mots. Ces expériences doivent néanmoins être refaites sur des données plus couramment utilisées dans le domaine (par exemple la collection Reuters).

3.3 Le multilinguisme

3.4 Importance du multilinguisme sur Internet

Les deux figures qui suivent montrent la répartition linguistique de la population Internet comparée à la population mondiale. La première représente cette répartition pour l'année 1999 (Loupy, 2000) (avec en plus la répartition par nombre de pages) et la seconde pour l'année 2000. On voit bien que, si la prédominance de l'anglais est nette, elle est en diminution sensible par rapport à d'autres langues. Et même les langues qui sont très présentes (allemand, espagnol, français, japonais) voient leur proportion décroître.

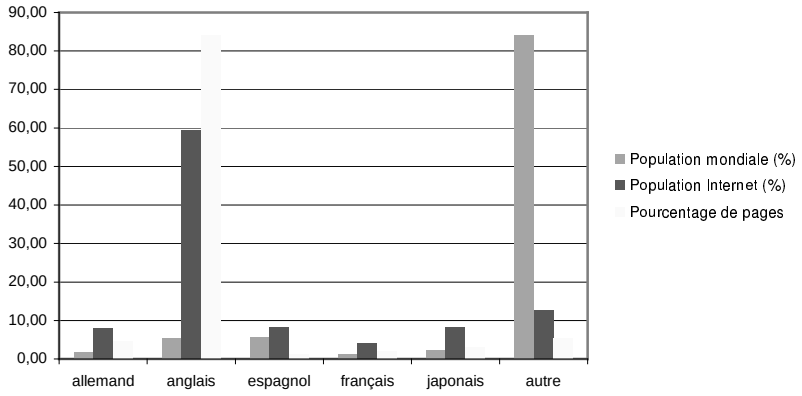

Figure 5 : La répartition des langues sur Internet et dans le Monde en 1999.

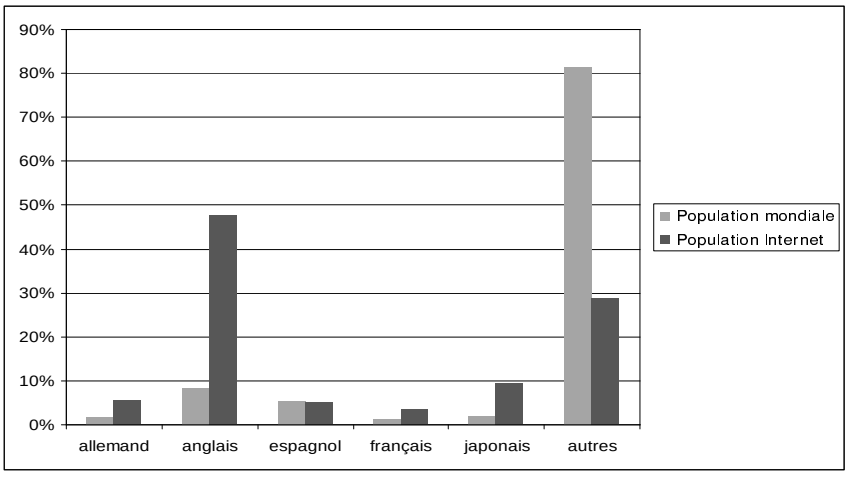

Figure 6 : La répartition des langues sur Internet et dans le Monde en 2000.

Tous les systèmes de recherche sur Internet sont confrontés à cet univers plurilingue. Or la gestion de plusieurs langues en parallèle pose de nombreux problèmes supplémentaires par rapport à une recherche monolingue. Dans le cas de la recherche inter-langue (requête en français, textes retrouvés en anglais), des ressources linguistiques sont nécessaires afin de connaître les traductions possibles d'un mot donné.

4 Connaissances linguistiques et interaction avec l'utilisateur

4.1 Utilité

Certains systèmes de recherche procèdent à une interaction avec l'utilisateur. Mais la quasi-totalité de ces interactions vise à enrichir la requête à l'aide de termes (synonymes ou provenant d'une extraction des premiers documents rapportés) en demandant une validation manuelle des termes en question. Ici, nous envisageons l'interaction sous un autre aspect.

Si nous pensons que l'utilisation de traitements ou de connaissances linguistiques apporte un avantage certain aux systèmes en matière de précision, il nous semble aussi que cela implique de prévoir une interaction avec l'utilisateur. En effet, le problème principal dans le traitement des requêtes est qu'elles sont en général très courtes (de 1 à 3 mots). Il est donc difficile de produire des traitements automatiques basés sur des règles syntaxiques par exemple. Mais, afin de tirer le meilleur parti des traitements linguistiques effectués lors de l'indexation, il serait préférable d'avoir un niveau d'analyse équivalent sur les requêtes. La connaissance de certaines caractéristiques linguistiques des termes utilisés dans la requête permet d'identifier les ambiguïtés ou les difficultés propres à la requête elle-même et donc d'en avertir l'utilisateur afin qu'il précise certains point ou qu'il reformule sa requête.

Ce genre d'interaction est utilisé depuis longtemps dans les systèmes d'aide à la traduction. Martin Kay donnait déjà, en 1973, le comportement d'un système de traduction idéal (Kay, 1973). Si la phrase à traduire était « They filled the tank with gas », le système pouvait demander : « Does the word 'tank' refer to : 1. a military vehicle, 2. a vessel for fluids ? ». Après réponse de l'utilisateur, le système était alors à même de traduire le mot ambigu. Bien que ce genre d'interaction ait été longtemps négligée, les besoins en recherche documentaire sont du même type.

4.2 Exemples de problèmes nécessitant une interaction

Il paraît utile de demander plus des précisions à l'utilisateurs dans de nombreux cas :

- Il est bien connu, depuis Luhn (1958), que plus un terme est fréquent, moins il apporte d'information pour un système de recherche documentaire. Si une requête n'est constituée que de termes fréquents il sera difficile d'y répondre correctement.

- Même si les mots de la requête sont de fréquence moyenne, si aucun texte ne les contient tous (pour des requêtes courtes), il risque d'être difficile de retrouver les documents pertinents.

- En plus de la présence de plusieurs termes de la requête, il faut pouvoir évaluer s'ils apparaissent à proximité l'un de l'autre. En effet, *organized* et *crime* sont très fréquents dans la collection TREC-6, mais le mot composé *organized crime* est beaucoup plus rare et donc plus précis. Il faut aussi tenir compte de ce phénomène.

- Si un terme a beaucoup de sens, le bruit risque d'être élevé. Une question visant à lever l'ambiguïté paraît utile.

- Le nombre de synonymes : si un terme (en fait un mot avec un sens donné) a beaucoup de synonymes, il est plus difficile de récupérer tous les documents pertinents par rapport au concept pointé, puisque beaucoup d'entre eux risquent d'utiliser des synonymes du terme de la requête pour représenter le concept.

- La dispersion des réponses dans des thématiques données ou extraites automatiquement est importante. Si les réponses ne montrent aucune cohérence dans les thématiques évoquées par les documents, il est probable que beaucoup de bruit a été généré.

5 Intérêt de la linguistique en recherche documentaire

Dans les sections précédentes, nous avons utilisé le point de vue classique, représenté par les évaluations TREC. Mais il faut noter que le but ultime d'une recherche documentaire est de répondre à un besoin informationnel de l'utilisateur. Si le système cherche des documents, l'utilisateur cherche presque toujours de l'information. La plupart des publications concernant la recherche documentaire ne basent leurs conclusions que sur des mesures de rappel et de précision. Or, il existe un grand nombre d'autres points qu'il conviendrait d'évaluer pour comparer deux systèmes de RD : couverture, rappel, précision, temps de réponse, effort fourni par l'utilisateur, présentation du résultat, taille de l'index, etc. Mais, par dessus tout, la rapidité de la recherche elle-même prise dans sa globalité semble fondamentalement importante. La figure suivante montre une évaluation prenant en compte le temps.

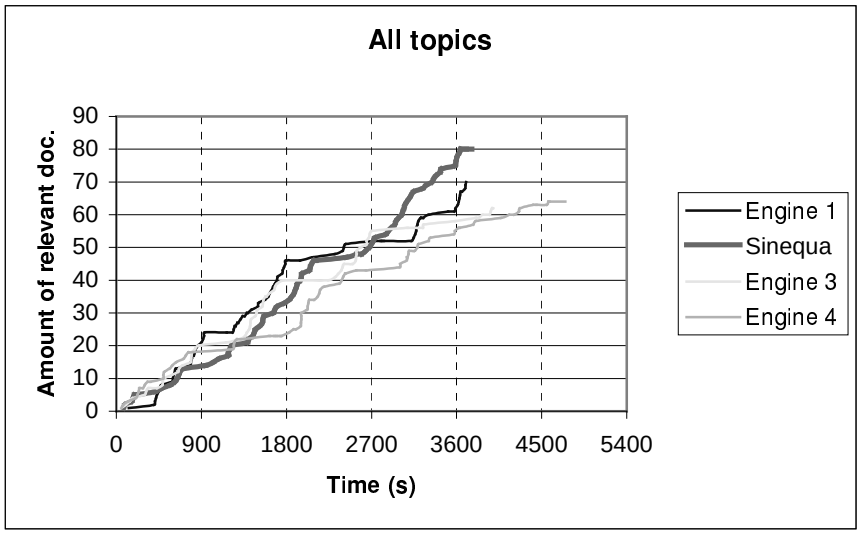

Figure 7 : Courbes d'évaluation de moteurs de recherche par Thalès

Les besoins principaux de l'utilisateur peuvent se résumer en : « trouver l'information voulue le plus rapidement possible ». L'efficacité du moteur de recherche doit donc être évaluée en tant qu'efficacité de la recherche de l'information par l'utilisateur et donc en faisant intervenir le temps de recherche. Thalès a récemment fait une étude de 4 moteurs de recherche (en prenant en compte le temps passé par les utilisateurs). Des courbes peuvent alors être tracées (figure 7) et le système de Sinequa prend rapidement la tête de l'évaluation.

Si l'on prend en compte le temps de recherche, la précision devient l'élément indispensable. En effet, un utilisateur ne regardera pas les mille premiers documents rapportés comme cela est évalué dans TREC, mais seulement quelques documents. Ainsi, Abondance publie un autre sondage dans lequel les sondés répondent à la question « Sur un outil de recherche (moteur, annuaire), combien de pages de résultats consultez-vous au maximum ? » (Abondance, 2000) :

1 page	4%
2 pages	24%
3 pages	23%
4 pages et plus	49%

Figure 8 : Sondage Abondance sur le nombre de documents consultés après une recherche.

Plus de 50 % des utilisateurs regardent de 1 à 3 pages au maximum. Il est donc nécessaire de placer les documents pertinents dès le début de la liste renvoyée. Or, nous avons vu que l'utilisation de connaissances et de traitements linguistiques permet d'obtenir une précision plus grande en tête de liste.

De plus, de nombreux moteurs fonctionnent en utilisant un système de retour de pertinence, avec validation de l'utilisateur ou non. Dans ce dernier cas, les systèmes traitent les premiers documents rapportés par le système afin d'extraire des termes clés qui serviront à enrichir la requête originelle. Pour ces systèmes, il est donc nécessaire que les premiers documents soient le plus pertinents possibles. Cela renforce encore le fait que l'utilisation de connaissances linguistiques permet d'améliorer les performances de ces outils.

6 Conclusion

Dans ce tutoriel, nous avons vu un certain nombre de difficultés liées à la langue et qui doivent être traitées par les systèmes de recherche documentaire. Les expériences publiées dans la littérature ne font pas apparaître clairement que les systèmes utilisant des connaissances linguistiques obtiennent de meilleures performances. Mais nous pensons que c'est parce que les éléments évalués ne sont pas les bons et qu'ils n'ont pas été évalués dans les bonnes conditions. Ainsi, les articles rapportant des expériences TREC se basent principalement sur la précision moyenne. Or, c'est la précision pour les premiers documents rapportés qui est la plus importante. De plus, le temps de recherche passé par l'utilisateur n'est pas pris en compte alors que c'est probablement l'élément le plus important dans la plupart des cas.

Dans ce contexte, nous avons montré que l'utilisation de connaissances et de traitements linguistiques permet d'augmenter la précision en tête de liste. Le fait que les premiers documents rapportés soient plus pertinents permet à l'utilisateur de retrouver plus vite

l'information qu'il cherche et aux systèmes utilisant un retour de pertinence automatique d'être plus performants.

Il ne s'agit pas, ici, d'opposer les moteurs utilisant des traitements linguistiques à ceux utilisant des méthodes statistiques. Tout moteur de recherche, sauf les plus primitifs, utilisent des méthodes statistiques pour faire ressortir les documents les plus pertinents. Les méthodes et les algorithmes sont connus depuis longtemps. Ce qui est souligné ici c'est qu'il est possible d'améliorer les performances des systèmes en leur ajoutant une couche linguistique.

Le couplage entre les outils statistiques et les connaissances linguistiques est, selon nous, le moyen le plus performant de satisfaire un utilisateur qui cherche à trouver au plus vite l'information dont il a besoin. Il est évident que le développement de ressources linguistiques et d'outils utilisant ces ressources est coûteux. Les systèmes utilisant des connaissances de ce type ne sont donc pas encore démocratisés. Mais il est probable qu'ils seront de plus en plus nombreux à l'avenir.

Enfin, les traitements linguistiques sont basés sur des règles relativement strictes d'utilisation des mots dans une langue donnée. Or, le contenu des documents que l'on trouve sur Internet par exemple ou dans des messages électroniques est souvent peu « linguistiquement correct ». Il est donc nécessaire de conserver le meilleur des deux mondes : des traitements simples et purement statistiques et des traitements fondés sur une connaissance linguistique forte.

Références

Abondance (2000), http://www.abondance.com/docs/sondages.html.

Bruce Croft W., Cook R., Wilder D. (1995), Providing government information on the internet : experiences with THOMAS, *Digital Libraries Conference DL'95*, pp. 19-24.

Fox c. (1990), A stoplist for general text, *SIGIR Forum*, 24 (1-2), pp. 19-35.

Gonzalo J., Verdejo F., Peters C., Calzolari N. (1998), Applying EuroWordNet to Cross-Language Text Retrieval, *Computers and the humanities, Special Issue on EuroWordNet*

Harman D. (1993), Overview of the First Text REtrieval Conference, *National Institute of Standards and Technology Special Publication 500-207.*

Hull D.A., Grefenstette G., Schultze B.M., Gaussier E., Schütze H., J.O. Pedersen (1996), Xerox TREC-5 site report: routing, filtering, NLP and Spanish tracks, *5th Text REtrieval Conference.*

Kay M. (1973), The MIND system, *Courant Computer Science Symposium 8: Natural language Processing*. Algorithmics Press, Inc. New York, pp. 155-188.

Kilgarriff A., Palmer M., (Editeurs) (2000) *Special Issue on SENSEVAL*, Computers and the Humanities.

Krovetz R. (1993), Viewing Morphology as an Inference Process, *16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 191-203.

Losee, R. M. (1996), *How part-of-speech tags affect text retrieval and filtering performance.*

Loupy C. de, Bellot P. (2000), Evaluation of document retrieval systems and query difficulty, *Using Evaluation within HLT Programs : Results and Trends*, pp. 34-40.

Loupy C. de, El-Bèze M. (2000), Using few cues can compensate the small amount of resources available for WSD, *Second International Conference on Language Resources and Evaluation.*

Loupy C. de, Bellot P., El-Bèze M., Marteau P.F. (1998a), Query expansion and classification of retrieved documents, *Seventh Text Retrieval Conference* (TREC-7), pp. 443-450.

Loupy C. de, El-Bèze M., Marteau P.-F. (1998b), Word Sense Disambiguation using HMM Tagger, *First International Conference on Language Resources & Evaluation*, pp. 1255-1258.

Loupy C. de (1999), Le codage des caractères dans les normes du document numérique, *Solaris*, No 6, <http://www.info.unicaen.fr/bnum/jelec/Solaris/d06/6loupy.html>.

Loupy C. de (2000), *Évaluation de l'apport de connaissances sémantiques en désambiguïsation sémantique et recherche documentaire*, Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse.

Luhn H.P. (1958), The automatic creation of litterature abstracts, *IBM Journal of Research and Development*, 2, pp. 159-165.

Palmer M. (1998), Are WordNet sense distinctions appropriate for computational lexicons?, *SENSEVAL Workshop*.

Royauté J., Schmidt L., Olivetan E. (1992), Les expériences d'indexation à l'INIST ; *COLING-92*, pp. 1058-1063.

Sanderson M. (1994), Word sense disambiguiation and information retrieval ; *17th annual international ACM-SIGIR conference on Research and development in information retrieval*, pp. 142-151.

Spriet T, El-Bèze M. (1997), Introduction of rules into a stochastic approach for language modelling, *Computational Models for Speech Pattern Processing, NATO ASI Series F*, editor K.M. Ponting.

Spriet T., Béchet F., El-Bèze M., Loupy C. de, Khouri L. (1996), Traitement automatique des mots inconnus ; *Troisième Conférence Annuelle sur le Traitement Automatique du Langage naturel*, TALN'96 ; pp. 170-179.

Wilks Y., Stevenson M. (1996), *The grammar of sense : is word-sense tagging much more than part-of-speech tagging ?*, Report No. CS-96-05, University of Sheffield.