Publié dans les
Actes 3ème Conférence en Recherche d'Information et Applications CORIA'06,
73-84, 2006
qui doivent être utilisés pour toute référence à ce travail

Un regard statistique sur l'évaluation de performance : L'exemple de CLEF 2005

Jacques Savoy

Institut interfacultaire d'informatique
Université de Neuchâtel, Pierre-à-Mazel 7, 2000 Neuchâtel (Suisse)
Jacques.Savoy@unine.ch

RÉSUMÉ. *Cette communication évalue et compare l'efficacité du dépistage de l'information de onze modèles à l'aide de quatre collections de documents rédigés dans les langues française, portugaise - brésilienne, hongroise et bulgare. Pour les deux dernières langues, on compare également l'indexation basée sur des mots à celle reposant sur des quadrigrammes (4-grams). En recourant à quatre tests statistiques et deux règles ad hoc, nous analysons les performances obtenues pour savoir si les différences de performance observées sont significatives. Enfin, nous comparons les résultats de ces différentes règles de décision afin de vérifier leur degré de concordance.*

ABSTRACT. *This paper evaluates and compares the retrieval effectiveness of eleven search models applied to four test collections written in the French, Portuguese-Brazilian, Hungarian and Bulgarian languages. For the latter two languages, we also compare word-based and 4-gram indexing schemes. Applying four statistical tests and two ad hoc rules, we analyze the performance levels obtained in order to determine whether their observed mean average performance differences are in fact statistically significant. Finally, we compare the results of these various decision rules and verify their degree of agreement.*

MOTS-CLÉS : *Recherche plurilingue ; évaluation ; tests statistiques.*

KEY WORDS: *Multilingual search; evaluation; statistical tests.*

1. Introduction

La recherche d'information repose, en partie, sur une tradition empirique. Dans ce cadre et au moyen de collections-tests, on désire savoir si une stratégie de dépistage s'avère réellement meilleure qu'une autre. Cette démarche se focalise essentiellement sur la capacité du système à extraire uniquement les documents pertinents en recourant aux mesures de précision (pourcentage de documents dépistés et pertinents par rapport à l'ensemble des documents extraits) ou de rappel (pourcentage de documents dépistés et pertinents par rapport à l'ensemble des documents pertinents). D'autres mesures de performance, telles que le temps de

réponse, *l'utilisabilité* du système de recherche ou la qualité de l'interface ne sont que peu étudiés, de manière empirique pour le moins. En effet, les études sur le comportement des usagers sont difficiles à élaborer et coûteuses à entreprendre. De plus, il s'avère souvent ardu d'isoler complètement l'effet du modèle de recherche de l'apport de l'interface homme - machine.

Si nous nous limitons aux mesures de précision et de rappel, nous devons reconnaître que de faibles différences ou un très faible nombre de requêtes ne sont pas suffisants afin d'aboutir à une conclusion définitive. En fait, une telle décision devrait se baser sur un test statistique démontrant qu'une différence de performance entre les deux systèmes n'a qu'une très faible chance d'être le fruit du hasard. Or dans un article récent [SAN 05], Sanderson & Zobel indiquent que l'usage de tel test n'est pas très répandu d'une part et, d'autre part, que certains auteurs semblent conclure assez rapidement à une différence significative même en présence d'une différence de quelques pourcents et en l'absence de test statistique.

Dans le cadre de cette communication, nous nous sommes intéressés à la piste *ad hoc* de la dernière campagne CLEF 2005 en présentant, dans la deuxième section, onze stratégies de recherche appliquées à une indexation par mot ou par quadrigramme. La section 3 présente brièvement quatre tests statistiques et deux règles *ad hoc* et les applique à nos expériences afin de savoir si leurs conclusions sont similaires voire identiques ou, au contraire, divergentes.

2. Recherche plurilingue

Nous désirons promouvoir la recherche d'information dans d'autres langues que l'anglais et, en particulier, en étudiant des langues proches (comme le français et le portugais) ou appartenant à la famille indo-européenne (comme le bulgare et son écriture cyrillique). Désirant tirer des conclusions très générales, nous inclurons également le hongrois (langue appartenant à la famille ouralienne) dans nos évaluations basées sur des corpus de la campagne d'évaluation CLEF-2005. Dans ce but, la section 2.1 décrit brièvement les quatre corpus utilisés dans nos expériences. Ensuite, la section 2.2 présente les modèles de dépistage de l'information que nous avons utilisés. Enfin, la section 2.3 évalue les modèles retenus pour les quatre langues en recourant à la précision moyenne, mesure officielle de performance des campagnes d'évaluation CLEF [PET 05].

2.1. Les collections-tests

Basé en partie sur notre participation à la campagne CLEF 2005, nos travaux portent sur l'interrogation unilingue de quatre corpus comprenant des articles de journaux ou d'agence de presse comme *Le Monde* (1994-1995, français), l'*Agence Télégraphique Suisse* (1994-1995, français), *Público* (1994-1995, portugais), *Folha* (1994-1995, brésilien), *Sega* (2002, bulgare), *Standart* (2002, bulgare) et *Magyar Hirlap* (2002, hongrois). Dans nos traitements automatiques, aucune distinction n'a

été faite entre le portugais et le brésilien. La table 1 indique quelques statistiques sur ces quatre corpus.

	Français	Portugais	Bulgare	Hongrois
taille	487 MB	564 MB	213 MB	105 MB
documents	177 452	210 734	69 195	49 530
termes / docum.	178	212,9	133,7	142,1
médiane	126	171	88	95
requêtes	50	50	49	50
nb doc. pertinent	2 537	2 904	778	939
pertinent / req.	50,74	58,08	15,88	18,78
maximum	185 (Q# 253)	239 (Q# 286)	69 (Q# 295)	87 (Q# 290)
minimum	1 (Q# 255)	2 (Q# 258)	1 (Q# 258)	1 (Q# 272)

Table 1 : *Quelques statistiques sur les quatre corpus*

Les besoins d'information exprimés couvrent des sujets divers (“Blanchiment d'argent”, ou “Blessures au football”), touchant parfois des sujets plutôt nationaux voire régionaux (“Référendums en Suisse”) ou, inversement, des thèmes possédant une couverture internationale (“L'euthanasie”). Comme pour d'autres campagnes d'évaluation, les requêtes se subdivisent en différents champs comme le titre (T) exprimant brièvement le thème de la requête, la partie descriptive (D) indiquant par une phrase le besoin d'information de l'usager et, finalement, la partie narrative (N) précisant les critères de pertinence.

2.2. Les stratégies d'indexation et les modèles de dépistage

Nous désirons obtenir une vision assez large de la performance de divers modèles de dépistage de l'information afin de pouvoir fonder nos conclusions sur des bases plus solides. Dans ce but, nous pouvons indexer les documents (et les requêtes) par un ensemble de termes sans aucune pondération (modèle noté “document=bnn, requête=bnn” ou “bnn-bnn”). Des modèles de dépistage plus performants ont été proposés dans lesquels l'importance de chaque terme retenu (dans un document ou une requête) tient compte de la fréquence d'occurrence (ou fréquence lexicale notée tf_{ij} pour le j^e terme dans le i^e document et le modèle correspondant se notera “nnn-nnn“). On peut également tenir compte de la fréquence documentaire d'un terme (ou df_j) ou plus précisément de l’$idf_j = \log(n/df_j)$. Chaque pondération peut encore être normalisé par le cosinus (modèle classique *tf idf* ou “ntc-ntc“). D'autres variantes dont les pondérations exactes sont reprises en annexe, ont été proposées, par exemple, pour imposer que la première occurrence d'un terme doit posséder plus d'influence (modèle “ltc“ ou “ltn“) ou dans lesquels la longueur du document jouera un rôle non négligeable (modèle “Lnu“ [BUC 96] ou “dtu“). Pour mesurer la similarité entre les documents et la requête,

on a utilisé le produit interne, soit pour le i^e document $\sum_{j=1,...t} w_{qj} \cdot w_{ij}$, avec w_{ij} indiquant le poids du j^e terme dans le i^e document et w_{qj} dans la requête.

En plus de ces solutions basées sur la vision géométrique du modèle vectoriel, nous avons considéré deux modèles probabilistes, à savoir l'approche Okapi [ROB 00] et le modèle GL2, un des membres de la famille *Divergence from ramdomness* [AMA 02]. Dans ce dernier cas, la pondération w_{ij} combine deux mesures d'information, à savoir :

$$\begin{aligned} w_{ij} &= Inf^1_{ij} \cdot Inf^2_{ij} = (1 - Prob^1_{ij}) \cdot -\log_2[Prob^2_{ij}] \\ Prob^1_{ij} &= tfn_{ij} / (tfn_{ij} + 1) \quad \text{avec } tfn_{ij} = tf_{ij} \cdot \log_2[1 + ((C \cdot mean\ dl) / l_i)] \\ Prob^2_{ij} &= [1 / (1+\lambda_j)] \cdot [\lambda_j / (1+\lambda_j)]^{tfn_{ij}} \quad \text{avec } \lambda_j = tc_j / n \end{aligned} \qquad (1)$$

dans laquelle l_i indique le nombre de terme d'indexation inclus dans la représentation du i^e document et tc_j représente le nombre d'occurrences du j^e terme dans la collection. Dans nos expériences, la constante *C* a été fixée de manière empirique à 1,25 (cette valeur offrait une légère amélioration par rapport à 1) et *mean dl* = 182 (FR), *mean dl* = 250 (PT), *mean dl* = 150 (HU) et *mean dl* = 134 (BG), soit des valeurs proches des statistiques de la table 1.

Pour ces langues européennes, nous proposons d'utiliser le mot comme unité d'indexation. Certes, les mots les plus fréquents ou appartenant à une forme grammaticale peu intéressante (conjonction, préposition, pronom, déterminant) sont éliminés. De même, nous procédons à la suppression automatique des suffixes liés à la flexion (pluriel, féminin, ainsi qu'à divers cas grammaticaux comme le génitif, l'ablatif, etc.) [SAV 06]. Comme alternative, nous avons indexé les collections hongroise et bulgare par des quadrigrammes (4-*grams*). Dans ce cas, on décompose chaque mot en séquence de quatre caractères contigus comme, par exemple, “jardin“ qui génère les trois formes “jard“, “ardi“ et “rdin“. Cette stratégie présente une bonne performance spécialement pour la langue hongroise [PET 05].

2.3. Evaluation selon la précision moyenne

Afin de mesurer la performance de ces différents modèles de dépistage, nous avons utilisé la précision moyenne à 11 points fixes de rappel (calculée par le logiciel `trec_eval` sur la base des 1 000 premières réponses). Cette mesure a été adoptée par les diverses campagnes d'évaluation pour évaluer la qualité de la réponse à des interrogations en ligne.

On constate que les modèles probabilistes Okapi ou GL2 présentent sur les quatre langues et, pour les deux représentations des documents (mot ou 4-*gram*), la meilleure performance. Par rapport à l'état de nos connaissances au début des années 90 avec le modèle *tf idf* (ou “ntc-ntc“ dans la table 2), ces modèles

proposent, en moyenne, une augmentation d'environ 35 % de la qualité de réponse. La différence entre les deux modèles probabilistes demeure faible (environ 1,5 %).

	Précision moyenne (requête TD)					
\ langue	Français	Portugais	Hongrois		Bulgare	
\ index	mot	mot	mot	4-*gram*	mot	4-*gram*
Okapi-npn	**0,3754**	**0,3477**	**0,3513**	**0,3914**	0,2706*	**0,2923**
GL2	0,3696*	0,3438*	0,3445	0,3877	**0,3030**	0,2868
Lnu-ltc	0,3437	0,3338*	0,3301	0,3545*	0,2737	0,2906
dtu-dtn	0,3365	0,3221*	0,3401	0,3409	0,2575*	0,2755
atn-ntc	0,3328	0,3076*	0,3215*	0,3498	0,2618	0,2301
ltn-ntc	0,3066	0,2535	0,2853	0,3139	0,2031	0,1433
lnc-ltc	0,2616	0,2535	0,2395	0,2725	0,2569	0,2674
ltc-ltc	0,2363	0,2234	0,2484	0,2809	0,2343	0,2393
ntc-ntc	0,2175	0,1868	0,2208	0,2767	0,1967	0,2110
bnn-bnn	0,0937	0,1322	0,1424	0,1075	0,0687	0,0458
nnn-nnn	0,0987	0,0639	0,0875	0,0576	0,0604	0,0144

***Table 2** : Précision moyenne de nos divers modèles de dépistage (indexation par mot ou quadrigramme (4-*gram*))*

Pour la collection hongroise, on constate que l'indexation basée sur les quadrigrammes (ou 4-*grams*) permet d'obtenir généralement une meilleure performance que celle reposant sur les mots (les modèles "bnn-bnn" et "nnn-nnn" sont les deux exceptions à cette règle). Pour la langue bulgare, les différences entre indexation par mot ou quadrigramme s'avèrent plus faibles d'une part et, d'autre part, aucune des deux formes d'indexation ne s'avère systématiquement meilleure que l'autre. Une explication du succès de l'indexation quadrigramme pour le corpus hongrois tient en partie à la présence de mots composés dans cette langue. Par exemple, la requête n° 255 contient le terme "internetfüggök" ("dépendance à internet") tandis que des variantes de cette forme apparaissent dans les documents pertinents ("internetfüggöséggel", "internetfüggöség" ou "internetfüggöségben"). Sur la base des mots, l'appariement échoue tandis que plusieurs quadrigrammes occurrent conjointement dans la requête et les documents pertinents.

Les valeurs de performance indiquées dans la table 2 indique que le modèle Okapi est meilleur que le modèle GL2 dans tous les cas sauf pour l'indexation par mots en bulgare. Cette conclusion demeurera-t-elle identique si nous ajoutons quelques requêtes ? Ou, pour le moins, peut-on inférer que les modèles probabilistes Okapi et GL2 proposent une performance toujours supérieure aux approches vectorielles ? La prochaine section abordera ces questions.

3. Tests statistiques

On ne saurait avancer une affirmation sur la base d'une seule observation et nous souhaiterions donc posséder un très grand nombre de requêtes afin de fonder nos conclusions. Mais l'établissement des jugements de pertinences correspondants représenteraient alors un coût prohibitif. Actuellement, les campagnes d'évaluation proposent de débuter avec un jeu de 50 requêtes, nombre jugé suffisant pour fonder des conclusions [VOO 02, SAN 05]. De plus, le jeu de requêtes disponibles doit représenter assez fidèlement les interrogations des utilisateurs. Cette dernière condition semble respectée ; il n'y a pas un nombre excessif de demandes liées à phénomène précis.

Notre méthodologie d'évaluation implique que chaque document pertinent possède la même valeur informative. Ainsi que l'utilisateur reçoive le premier document pertinent ou le cinquième, la valeur ajoutée sera identique. Ce choix ne reflète pas complètement la réalité mais facilite les mesures de précision et de rappel. Cette hypothèse d'utilité marginale constante est très rare en modélisation économique. En effet, pour un bien donné, il est reconnu que l'utilité diminue avec la quantité. Ainsi, si j'ai bien apprécié mon premier croissant ce matin, le deuxième avait déjà une utilité moins élevée, tandis que le troisième possédait une faible utilité.

Sur ces hypothèses méthodologiques, nous désirons savoir si les différences de performance sont statistiquement significatives entre le meilleur modèle et les autres. Plus particulièrement, sachant que plusieurs tests statistiques ont été proposés pour répondre à cette interrogation, est-ce que leurs conclusions convergent ? De plus, est-ce que l'approche basée sur le ré-échantillonnage aléatoire (*bootstrap*) [SAV 97] possède un comportement différent des autres tests statistiques classiques ? Enfin, deux règles *ad hoc* simples ont été proposées afin de déterminer si la différence entre deux approches peut être qualifiée de significative [SAN 05]. Dans ce dernier cas, est-ce que les décisions qui en découlent sont identiques à celles issues d'un test statistique ? Afin de répondre à ces questions, nous présenterons brièvement quatre tests statistiques dans la section 3.1 puis, dans la section 3.2, nous appliquerons ces règles de décision aux performances indiquées dans la table 2 en recourant au logiciel R. Enfin, dans la dernière section, nous comparerons l'indexation par mot à celle basée sur les quadrigrammes pour les langues hongroise et bulgare.

3.1. Principes des tests proposés

En recherche d'information, trois principaux tests statistiques sont utilisés pour détecter une différence significative entre deux précisions moyennes. Ces trois approches reposent sur les principes communs suivants. On évalue les deux systèmes sous les mêmes conditions (requêtes identiques) et on admet que le

traitement d'une requête n'a aucune influence sur le traitement d'autres requêtes, que ce soit sur le même système ou sur le système concurrent (indépendance). On peut donc calculer, pour chaque requête, la différence de performance entre les deux modèles. Le traitement de ces différences diverge quelque peu selon le test utilisé.

D'abord, on peut recourir au test de Student pour observations pairées [GRI 93]. Ce test suppose que la distribution des différences de performance suive une loi normale, ce qui implique que la distribution soit symétrique et que son mode, sa médiane et sa moyenne correspondent à la même valeur. Dans nos expériences, sur 82 distributions analysées, 11 s'écartaient significativement d'une loi normale selon le test de Kolmogorov (seuil de signification 5 %). On admet parfois que même si les conditions exigées par le test de Student ne sont pas formellement respectées, les conclusions de ce test demeurent valide [HUL 93], [SAN 05].

Refusant d'imposer ou d'admettre ces conditions, le test du signe [CAN 80] suppose que, pour chaque requête, on puisse affirmer que l'un des systèmes est meilleur que l'autre ou, au contraire, que les deux systèmes apportent une réponse de qualité identique. Pour le test des rangs signés de Wilcoxon [CAN 80], on impose que la distribution des différences est symétrique, sans requérir que cette distribution suive une loi précise. La différence majeure entre ces deux tests non-paramétriques est la suivante. Pour une requête donnée, le test du signe tient compte uniquement de l'information que l'un des systèmes est meilleur que l'autre. Par contre, pour le test de Wilcoxon, une différence plus importante aura aussi une influence plus forte dans la décision finale. Ainsi, une différence de performance de 0,01 ou de 0,2 possédera le même impact avec le test du signe[1] mais pas dans le test de Wilcoxon.

Comme alternative à ces tests, nous pouvons recourir au ré-échantillonnage aléatoire non-paramétrique (*bootstrap*) qui n'impose pas de contraintes spécifiques [SAV 97]. Cette démarche propose également un test permettant de vérifier si deux systèmes proposent la même performance moyenne.

Dans nos analyses, l'hypothèse nulle, nommée H_0, sera toujours la même à savoir que « les deux modèles de dépistage offrent la même performance moyenne ». Avec chaque test on peut alors calculer une valeur *p*, la probabilité d'obtenir dans l'échantillon les valeurs observées ou plus extrêmes sachant que H_0 est vraie. Si cette probabilité est inférieure à 0,05 (seuil de signification de notre test bilatéral), nous rejetterons H_0 au profit de l'hypothèse alternative (« il existe une différence de performance entre les deux systèmes »).

[1] Dans le cadre du test du signe, il n'est pas interdit d'imposer une différence minimale qui devra être dépassée afin de conclure que, pour cette requête, l'un des deux systèmes présente une meilleure qualité de réponse.

Finalement, on a également proposé deux règles de décision *ad hoc*. D'abord, on peut admettre qu'une différence absolue de 0,05 de précision moyenne (avec 50 requêtes ou plus) peut être jugée comme significative [SPA 77; VOO 02]. Ensuite, face à une différence relative plus élevée que 25 % (précision moyenne), on peut admettre que les deux approches proposent des performances distinctes [SAN 05].

3.2. Evaluation statistique

Sur 60 différences de performance analysées dans la table 2, les quatre tests aboutissent à la même conclusion dans 51 observations. Pour être plus précis, ces 51 cas se subdivisent en neuf cas pour lesquels les quatre tests indiquent qu'il n'y a pas de différence significative et, pour les 42 autres, soulignés dans la table 2, la différence est perçue comme significative (test bilatéral par rapport à la meilleure performance indiquée en gras, seuil à 5 %). Pour neuf observations restantes (signalées par un petit astérisque "*"), la conclusion des quatre tests diverge. La table 3 résume la situation pour quatre de ces observations.

Pour la collection française (deux premières lignes de la table 3), la différence entre Okapi (0,3754) et GL2 (0,3696) est très faible (différence relative de 1,5 %). Le modèle Okapi propose une meilleure performance pour 36 requêtes et GL2 pour 14. Pour le test du signe, la valeur *p* s'élève à 0,0026 et elle est considérée comme significative (car inférieure à 0,05). Pour le test de Student, la différence entre les moyennes est trop faible, compte tenu du nombre d'observations et de la variabilité de la performance entre requêtes, pour être détectée comme significative (valeur *p* est de 0,1795, supérieure à 0,05).

			valeur *p*			
Modèles	préc. moy.	# requêtes	Student	signe	Wilcoxon	*bootstrap*
FR / Okapi	0,3754	36				
GL2	0,3696	14	**0,1795**	0,0026	0,0239	**0,1719**
PT / Okapi	0,3477	31				
atn-ntc	0,3076	19	0,0008	**0,1189**	0,0015	0,0015
HU / Okapi	0,3513	33				
atn-ntc	0,3215	17	**0,0605**	0,0328	0,0169	0,0425
BG / GL2	0,3030	29				
dtu-dtn	0,2575	18	0,0182	**0,1439**	0,0214	0,0160

***Table 3** : Quelques comparaisons de systèmes avec des conclusions divergentes*

Pour la collection portugaise, la différence entre les modèles Okapi (0,3477) et "atn-ntc" (0,3076, soit une différence relative de 11,53 %) est significative pour trois tests. Par contre, pour le test du signe, la valeur *p* s'élève à 0,1189 car la différence entre les 31 "+" et 19 "-" n'est pas assez nette. Pour la collection hongroise

(indexation par mot), la différence relative entre Okapi (0,3513) et “atn-ntc“ (0,3215) s'élève à 8,48 %. Dans ce cas, le test de Student indique une valeur *p* de 0,0605 soit légèrement supérieure à notre limite de 0,05. Pour l'indexation par mot dans le corpus bulgare, trois tests statistiques indiquent que la différence entre le modèle GL2 (0,3030) et “dtu-dtn“(0,2575) est significative. Pour le test du signe, le nombre de requêtes favorisant le modèle GL2 (29 requêtes) n'est pas suffisant pour franchir la limite des 0,05 (valeur *p* de 0,1439).

Dans le contexte de nos évaluations, les règles *ad hoc* sont plus conservatrices que les quatre tests statistiques. Ainsi, si l'on considère comme significative les différences supérieures en valeur absolue à 0,05, on en dénombre seulement 35. Comme ces 35 comparaisons de performance forment un sous-ensemble parmi les 49 cas détectés par nos quatre tests, il n'y a pas de contraction importante. Avec les valeurs de précision moyenne indiquées dans la table 2, la règle *ad hoc* des 25 % indique que la différence sera perçue comme significative pour 29 cas seulement (formant un sous-ensemble des 35 observations de la règle des 0,05). Sur la base des valeurs maximales de la table 2, une différence de 25 % revient à tenir compte d'une différence absolue entre 0,075 (si le meilleur système possède une précision moyenne de 0,3) ou de 0,1 (avec une précision moyenne maximale de 0,4).

Finalement, on a remarqué que la plus petite différence significative, aux yeux de nos quatre tests statistiques, s'élevait à 0,0317 ou à 8,44 % (corpus français, Okapi (0,3754) vs. “Lnu-ltc“ (0,3437)). D'un autre côté, la différence de performance du modèle GL2 avec l'indexation par mot ou quadrigramme (corpus hongrois) propose la plus forte différence jugée non-significative par tous les tests statistiques (0,3513 vs. 0,3914, différence absolue de 0,0401). En pourcentage, la plus grande différence non-significative pour nos quatre tests s'élève à 13,78 % (collection bulgare, modèle “atn-ntc“ et comparaison entre indexation par mot et quadrigramme, soit 0,2618 vs. 0,2301).

3.3. Indexation par mot ou quadrigramme ?

Dans la table 2, nous avons comparé la précision moyenne obtenue avec une indexation s'appuyant sur les mots ou sur les quadrigrammes pour les langues hongroise et bulgare. Pour la langue bulgare, il n'y a pas vraiment de raison de privilégier une forme d'indexation au profit de l'autre. Pour les onze modèles, seul deux différences sont significatives au regard des quatre tests statistiques, à savoir pour les modèles “ltn-ntc“ (0,2031 vs. 0,1433) et “nnn-nnn“ (0,0604 vs. 0,0144). Dans ces deux cas, la meilleure approche repose sur une indexation par les mots.

Pour la collection hongroise, la situation décrite dans la table 2 favorisait une indexation par les quadrigrammes. Pourtant, aucune des différences n'est perçue comme significative pour les quatre tests. Si a priori on peut soutenir que l'indexation par quadrigramme est meilleure que l'indexation par mot, au moins pour

le hongrois, cette conclusion n'est simplement pas confirmée par les tests statistiques. Cependant, la différence de performance avec le modèle "ntc-ntc" (soit 0,2208 - mot - et 0,2767 - 4-*gram* -) s'avère significative pour trois tests ; le test du signe possède une valeur *p* de 0,1524 qui est supérieure à notre limite de 5 %. De plus, on notera que le test du *bootstrap* détecte une différence significative avec le modèle GL2 (0,3445 - mot - et 0,3877 - 4-*gram* -, valeur $p = 0,0385$). Dans ce cas, le test de Student indique que la différence est presque significative (valeur $p = 0,0512$) tandis que pour le test du signe, la valeur *p* est nettement plus élevée (0,4799) car l'indexation par quadrigramme est meilleure pour 28 requêtes contre 22 pour l'indexation par mot.

Selon la règle *ad hoc* du 0,05, la différence est significative pour le modèle "ntc-ntc"(hongrois) et, selon la règle ad hoc des 25 %, les différences significatives sont celles apparaissant avec les approches "nnn-nnn" et "bnn-bnn". Pour le corpus bulgare, les deux règles *ad hoc* de décision indiquent aussi que la différence de performance doit être vue comme significative pour le modèle "ltn-ntc". Pour la règle des 25 %, on doit encore ajouter les systèmes "nnn-nnn" et "bnn-bnn" dont la performance avec l'indexation par mot est supérieure à celle des quadrigrammes de plus de 25 %. Ces exemples illustrent la limite des deux règles *ad hoc* ; dès que l'une des performances que l'on compare est très faible, il s'avère aisé de l'accroître pour dépasser le seuil de 25 % ou celui de 0,05.

4. Conclusion

Sur la base de quatre corpus extraits de CLEF 2005, nous avons démontré que les meilleures précisions moyennes s'obtiennent avec les modèles probabilistes Okapi ou GL2. De plus, il n'y a pas de raison objective de privilégier l'une des ces deux approches probabilistes. Pour le hongrois par exemple, il suffit d'éliminer deux requêtes pour inverser les deux premières places au classement. Si l'on compare l'indexation par mots ou par quadrigrammes, les performances sont assez similaires avec la collection bulgare. Par contre, pour la langue hongroise, les quadrigrammes semblent apporter une performance plus élevée. Comme cette approche était la meilleure lors de la campagne CLEF 2005, la conclusion qui en découle semble évidente. Or l'ensemble des quatre tests statistiques ne s'accordent pas sur cette affirmation, indiquant qu'il n'y a pas de raison objective de privilégier une indexation au profit de l'autre.

De plus, on constate que les quatre tests disponibles proposent des conclusions convergentes mais pas rigoureusement identiques. Ainsi, si deux valeurs de performance sont assez proches, le test de Student ne détectera pas une différence significative, surtout si l'on considère le nombre relativement faible d'observations (50) pour un tel cas de figure. Pour le test du signe par contre, la différence peut être analysée comme importance car elle se répète pour de nombreuses requêtes

(voir table 3). A l'inverse, si un modèle apporte une amélioration très sensible de la performance pour quelques requêtes, la moyenne en sera clairement augmentée. Dans ce cas, le test de Student (ou celui basé sur le *bootstrap*) signalera une différence significative que le test du signe ne confirmera peut-être pas.

On a également remarqué que les résultats du test de Student et ceux du ré-échantillonnage aléatoire débouchent très régulièrement sur la même décision. Les deux règles *ad hoc* (0,05 et 25 %) sont plus conservatrices que les tests statistiques. Cependant, de telles règles de décision devraient être utilisées avec précaution si l'une des valeurs à comparer s'avère très faible (par exemple, inférieure à 0,1). Finalement, signalons que si une différence qui n'est pas considérée comme significative par un test statistique, elle peut tout de même être perçue comme importante par les utilisateurs si elle se répète dans différents contextes.

Remerciements

Cette recherche a été financée en partie par le Fonds national suisse pour la recherche scientifique (subsides n^0 21-66 742.01 et n^0 200020-103420).

5. Bibliographie

[AMA 02] Amati, G., van Rijsbergen, C.J. "Probabilistic models of information retrieval based on measuring the divergence from randomness", ACM-Transactions on Information Systems, vol. 20, n° 4, 2002, p. 357-389.

[BUC 96] Buckley, C, Singhal, A., Mitra, M., Salton, G., "New retrieval approaches using SMART", Proceedings of TREC-4, NIST Publication #500-236, Gaithersburg (MD), 1996, p. 25-48.

[CON 80] Conover, W.J. Practical nonparametric statistics. 2nd ed., John Wiley & Sons, New York, 1980.

[GRI 93] Grimm, L.G. Statistical applications for the behavioral sciences. John Wiley & Sons, New York, 1993.

[HUL 93] Hull, D. "Using statistical testing in the evaluation of retrieval experiments", Proceedings of ACM-SIGIR'93, Pittsburgh (PA), 1993, p. 329-338.

[PET 05] Peters, C. Working Notes for the CLEF 2005 Workshop. 2005, see the URL http://www.clef-campaign.org/2005/working_notes/ (visité le 25 octobre 2005).

[ROB 00] Robertson, S.E., Walker, S., Beaulieu, M., "Experimentation as a way of life: Okapi at TREC", Information Processing & Management, vol. 36, n° 1, 2000, p. 95-108.

[SAN 05] Sanderson, M., Zobel, J. "Information retrieval system evaluation: Effort, sensitivity, and reliability", Proceedings of ACM-SIGIR 2005, The ACM Press, New York, 2005, p. 162-169.

[SAV 97] Savoy, J., "Statistical inference in retrieval effectiveness evaluation", Information Processing & Management, vol. 33, n° 4, 1997, p. 495-512.

[SAV 06] Savoy, J., Berger, P.-Y., "Monolingual, bilingual and GIRT information retrieval at CLEF-2005", Proceedings CLEF 2005, Springer-Verlag, Berlin, 2006, to appear.

[SPA 77] Sparck Jones, K., Bates, R.G. Research on automatic indexing 1974-1976. Technical Report, Computer Laboratory, University of Cambridge (UK), 1977.

[VOO 98] Voorhees, E.M., Harman, D. "Overview of the sixth TExt Retrieval Conference (TREC-6)", Proceedings of TREC-6, NIST Publication #500-240, Gaithersburg (MD), 1998, p. 1-24.

[VOO 02] Voorhees, E.M., Buckley, C. "The effect of topic set size on retrieval error evaluation measure stability", Proceedings of ACM-SIGIR 2002, The ACM Press, New York, 2005, p. 316-323.

[VOO 04] Voorhees, E.M. "Overview of the TREC 2004 robust retrieval track", Proceedings of TREC-2004, NIST Publication, Gaithersburg (MD), 2004.

Annexe 1. Formules de pondération

Dans les formules décrites dans la table 4, *n* indique le nombre d'articles dans le corpus, *t* le nombre de termes d'indexation différents, tf_{ij} le nombre d'occurrences du j^e^ terme dans le i^e^ document, df_j le nombre d'articles indexés avec le j^e^ terme, avec $idf_j = \ln(n/df_j)$, nt_i indique la longueur (nombre de termes d'indexation distincts) du i^e^ article, et l_i le nombre de termes d'indexation du i^e^ document. La valeur des constantes a été fixée ainsi (de manière empirique, la performance étant la meilleure) : *b*=0,7 (FR et PT), *b*=0,75 (HU et BG), k_1=1,5, *avdl*=600 (FR), *avdl*=700 (PT), *avdl*=750 (HU et BG), *pivot*=100 et *slope*=0.2.

bnn	$w_{ij} = 1$	nnn	$w_{ij} = tf_{ij}$
ltn	$w_{ij} = [\ln(tf_{ij}) + 1] \cdot idf_j$	atn	$w_{ij} = \left[0,5 + 0,5 \cdot \frac{tf_{ij}}{\max tf_{i.}}\right] \cdot idf_j$
dtn	$w_{ij} = \ln[[\ln(tf_{ij}) +1] +1] \cdot idf_j$	npn	$w_{ij} = tf_{ij} \cdot \ln\left[\left(n - df_j\right) \Big/ df_j\right]$
Lnu	$w_{ij} = \frac{\left(1 + \ln(tf_{ij}) \Big/ \ln(\text{mean tf}) + 1\right)}{(1 - slope) \cdot pivot + slope \cdot nt_i}$	Okapi	$w_{ij} = \left((k_1 + 1) \cdot tf_{ij}\right) \Big/ \left(K + tf_{ij}\right)$
lnc	$w_{ij} = \frac{\ln(tf_{ij}) + 1}{\sqrt{\sum_{k=1}^{t} \left(\ln(tf_{ik}) + 1\right)^2}}$	ntc	$w_{ij} = \frac{tf_{ij} \cdot idf_j}{\sqrt{\sum_{k=1}^{t} \left(tf_{ik} \cdot idf_k\right)^2}}$
dtu	$w_{ij} = \frac{\left[\ln\left(\ln(tf_{ij}) + 1\right) + 1\right] \cdot idf_j}{(1 - slope) \cdot pivot + slope \cdot nt_i}$	ltc	$w_{ij} = \frac{\left(\ln(tf_{ij}) + 1\right) \cdot idf_j}{\sqrt{\sum_{k=1}^{t} \left(\left(\ln(tf_{ik}) + 1\right) \cdot idf_k\right)^2}}$

Table 4 : *Formules de pondération utilisées*