Publié dans les
Actes 6ème Conférence en Recherche d'Information et Applications CORIA'09,
19-31, 2009
qui doivent être utilisés pour toute référence à ce travail


# Evaluation de diverses stratégies de désambiguïsation lexicale

**Claire Fautsch, Jacques Savoy**

*Institut d'informatique*
*Université de Neuchâtel, rue Emile Argand 11, 2009 Neuchâtel (Suisse)*
*Claire.Fautsch@unine.ch, Jacques.Savoy@unine.ch*

RESUME. *Dans la campagne d'évaluation CLEF-2008, la tâche « robuste » fournissait un corpus enrichi en langue anglaise. Pour chaque mot, le lemme, la partie du discours et le numéro Synsets de WordNet™ (numéro de classe d'un thésaurus) étaient fournis. Sur cette base, nous avons testé plusieurs approches afin de lever, en partie pour le moins, l'ambiguïté lexicale. Recourant au modèle vectoriel tf idf, ainsi qu'à trois approches probabilistes et un modèle de langue, cet article évalue leur performance en fonction de diverses techniques d'enracineur. Un enracineur léger permet d'obtenir des performances similaires à des approches plus agressives ou à celle obtenue par une analyse morphologique. L'indication de la partie du discours permet d'améliorer significativement la qualité de la réponse tandis que les numéros de classes d'un thésaurus n'ont pas permis une amélioration.*

ABSTRACT. *In the robust track of the 2008 CLEF evaluation campaign an enlarged English corpus was provided. For each term, the lemma, the part-of-speech (POS) and the Synset number extracted from WordNet™ (class number of the corresponding thesaurus) are given. Based on this corpus we tested several approaches to remove at least partially the underling lexical ambiguity. Using different IR models such as the vector-space model tf idf as well as three probabilistic models and a language model, we want to evaluate their performance when using different algorithmic or morphological stemming approaches. The inclusion of the part-of-speech information improves the retrieval performance significantly, while the inclusion of the synset number does not show any improvement.*

MOTS-CLES : *Analyse morphologique, enracineur, thésaurus, partie du discours, évaluation.*

KEY WORDS: *Morphological Analysis, Stemmer, Thesaurus, Part-Of-Speech, Evaluation.*

## 1. Introduction

En recherche d'information (RI), l'emploi d'enracineurs représente une technique fréquente qui, en général, améliore la qualité des résultats obtenus [MAN 08]. Le but recherché est de permettre des appariements entre des formes sémantiquement reliées mais dont l'orthographe varie. Par exemple, si une requête contient le mot "chat" et le document renferme la forme "chats", l'enracineur devrait regrouper ces deux formes de surface sous la même entrée dans l'index de même que pour les formes "chattes" ou "chaton".

Pour la langue anglaise, différents enracineurs algorithmiques ont été proposés basés sur des règles morphologiques de la langue comme, par exemple, celui de Porter [POR 80] ou Lovins [LOV 68]. Contrairement à ceux-ci, nous pouvons procéder à une analyse morphologique plus profonde, requérant certes plus de ressources (dictionnaire), mais permettant de retourner le lemme du mot (on son entrée dans le dictionnaire). Cette première étape nous permet d'éliminer toutes les marques liées aux flexions. De plus en se basant sur la partie du discours (POS) nous pourrions améliorer la qualité des appariements entre formes identiques (e.g., le sens de "*mean*" comme nom et comme verbe est différent) ou utiliser l'information POS pour mieux contrôler l'élimination des suffixes dérivationnels [SAV 93]. Comme approche supplémentaire pour améliorer le dépistage de documents pertinents, nous pourrions utiliser le numéro de la classe Synset du thésaurus WordNet™ [FEL 98] afin de faciliter l'appariement entre formes différentes mais ayant un sens proche.

Cet article a pour objectif d'évaluer et d'analyser l'impact de divers traitements morphologiques (enracineur, lemmatisation, partie du discours et thésaurus) disponible avec la tâche « robuste » de la campagne d'évaluation CLEF 2008. La suite de cette communication est organisée de la manière suivante. Dans la deuxième section nous décrivons brièvement le corpus utilisé tandis que la troisième section présente les enracineurs et les modèles de dépistage utilisés. Les évaluations faites sont exposées dans la quatrième section et une conclusion résume les principales contributions dans une cinquième section.

## 2. Regard sur le corpus d'évaluation

La tâche « robuste » de la campagne d'évaluation CLEF-2008 a décidé de former un large ensemble de requêtes en regroupant l'ensemble des collections rédigées en langue anglaise et couvrant les campagnes de 2001 à 2006 (Peters *et al*. 2008). Ce corpus se compose d'articles de journaux du *Los Angeles Times* publiés durant l'année 1994 ainsi que des documents du *Glasgow Herald* parus en 1995. Ce corpus comprend un total de 169 477 documents (correspondant à environ 579 MB de données). En moyenne, chaque article contient environ 250 mots pleins (médiane:

191) (ce calcul ne tient pas compte des mots outils comme “the”, “of” ou “in”). Un document caractéristique possède un titre bref suivi d'un à quatre paragraphes de texte pouvant être rédigés selon l'orthographe anglaise ou américaine.

La figure 1 illustre avec quelques détails les diverses composantes de notre corpus d'évaluation. Par exemple, pour l'année 2003, les requêtes disponibles débutent au numéro 141 et s'achève avec le numéro 200. Sur cet ensemble, nous disposons de 54 interrogations avec au moins un document pertinent. Ces bonnes réponses doivent être dépistées selon l’année dans un ou deux des journaux *Los Angeles Times* et *Glasgow Herald* comme indiquée dans la figure 1. Par exemple en 2004 on ne tient compte que du *Glasgow Herald* tandis qu’en 2005 on recherche l’information dans les deux journaux.

| | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|---|---|---|---|---|---|---|
| Source | *LA Times* | *LA Times* | *LA Times Glasgow H.* | *Glasgow H.* | *LA Times Glasgow H.* | *LA Times Glasgow H.* |
| Taille | 425 MB | 425 MB | 579 MB | 154 MB | 579 MB | 579 MB |
| # docs | 113 005 | 113 005 | 169 477 | 56 472 | 169 477 | 169 477 |
| Nb req. | 47 | 42 | 54 | 42 | 50 | 49 |
| Requête | n°41-n°90 | n°91-n°140 | n°141-n°200 | n°201-n°250 | n°251-n°300 | n°301-n°350 |

***Figure 1** : Caractéristiques essentielles des diverses parties de notre corpus d'évaluation*

Suivant le modèle des campagnes TREC, chaque requête possède principalement trois champs logiques, à savoir un titre bref (T), une phrase décrivant le besoin d’information (D) et une partie narrative (N). La figure 2 présente, dans sa partie supérieure, un exemple. Pour l’essentiel de nos évaluations, nous avons retenu uniquement la partie “titre” (T) pour construire les requêtes (pour, en moyenne, 2,91 termes d’indexation par requête) ou les parties “titre” et “descriptif” (TD) (moyenne, 7,51 termes par requête).

Lors de la campagne d’évaluation CLEF 2008, les organisateurs ont ajouté des informations tant au niveau des documents que des requêtes. La partie inférieure de la figure 2 illustre un exemple d’une requête élargie. Ainsi, avec chaque mot, on retrouve sa forme actuelle (sous l'étiquette <WF>), sa partie du discours (étiquette <POS>), son lemme (ou son entrée dans le dictionnaire avec l'étiquette <LEMA>) et finalement le ou les numéros de classes dans le thésaurus WordNet™ (étiquette <SYNSET>). L'ensemble de ces informations a été ajouté afin de mesurer l'efficacité de diverses stratégies pouvant éliminer ou, pour le moins réduire, l'ambiguïté lexicale en recherche d'information.

Ainsi avec le lemme fourni nous pouvons travailler directement avec le résultat d'une analyse morphologique. Le recours à un enracineur léger éliminant les flexions morphologiques s'avère superflu. La partie du discours peut fournir des indications pertinentes pour procéder à la suppression des suffixes de dérivation ou pour favoriser des appariements entre mots de même nature. Le nombre associé à l'entrée dans le thesaurus WordNet™ (version 1.6) permet de définir les synonymes d'un terme. Evidement pour les noms propres (nom de personne, lieu ou de produit comme "Haiti" ou "Kaurismäkis"), cette information n'existe pas. Dans le corpus ce nombre peut être unique si le terme n'apparait que dans une seule classe. Parfois plusieurs numéros de classe du thésaurus sont indiqués avec un score représentant la probabilité que le Synset correspondant soit correct (par exemple, dans la figure 2, deux numéros de classe sont attribués pour le mot "Bankruptcy").

```
<NUM> C180 </NUM>
<EN-TITLE> Bankruptcy of Barings </EN-TITLE>
<EN-DESC> What was the extent of the losses in the Barings bankruptcy case?
 </EN-DESC>
<EN-NARR> Relevant documents must quantify in some way the losses caused by
the collapse of the oldest bank in Great Britain </EN-NARR>
...
<NUM> C180 </NUM>
<EN-TITLE>
 <TERM ID="10.2452/180-AH-1" LEMA = "bankruptcy" POS = "NNP">
 <WF> Bankruptcy </WF>
  <SYNSET SCORE = "0.4819665883771086" CODE = "10386276-n"/>
  <SYNSET SCORE = "0.5180334116228914" CODE = "10386165-n"/> </TERM>
<TERM ID = "10.2452/180-AH-2" LEMA = "of" POS = "IN">
 <WF> of </WF> </TERM>
<TERM ID = "10.2452/180-AH-3" LEMA = "baring" POS = "NNPS">
 <WF> Barings </WF>
 <SYNSET SCORE = "1" CODE = "00819570-n"/> </TERM>
</EN-TITLE>
...
```

***Figure 2** : Exemple d'une requête avec les indications du lemme, de sa partie du discours, et des numéros de classe de WordNet associées*

L'ensemble des ces informations n'a pas été ajouté manuellement mais en recourant à différents traitements automatiques. En premier lieu, le système MXPOST (*Maximum Entropy POS Tagger*[1]) [RAT 96] a été utilisé afin de déterminer la partie du discours (POS) pour chaque terme. Lors d'une deuxième étape, le lemme correspondant est extrait de WordNet™ en utilisant JWNL (*Java WordNet Library*), une API permettant un accès facile au thésaurus. Finalement, en

[1] Téléchargeable sur http://www.inf.ed.ac.uk/resources/nlp/local_doc/MXPOST.html

se basant sur ces informations, des collocations locales et le contexte, le système de désambiguïsation NUS-PT [CHA 07] détermine la ou les classes du thésaurus correspondant au mot étudié. Cette affectation s'effectue, pour l'essentiel, avec un algorithme de type SVM (*Support Vector Machine*) entrainé sur différents corpus dont des collections d'articles de journaux, type d'information que l'on retrouve dans les collections CLEF. Ces divers traitements sont pas excepte d'erreurs comme le fait que l'étiquette POS associé au mot “Bankruptcy” dans la figure 2 est “NNP” (nom propre) et non “NN” (nom).

## 3. Modèles de recherche d'information

Afin d’obtenir une vision assez large de la performance de divers traitements lexicaux, nous avons retenus différentes approches. Comme modèle de base, nous avons indexé les documents (et les requêtes) selon la formulation classique $tf \cdot idf$ qui tient compte de la fréquence d'occurrence (ou fréquence lexicale notée $tf_{ij}$ pour le $j^{e}$ terme dans le $i^{e}$ document) et de la fréquence documentaire d'un terme ($df_j$, ou plus précisément de l’$idf_j = \log(n/df_j)$ avec $n$ indiquant le nombre de documents inclus dans le corpus).

L'évaluation avec ce modèle vectoriel sera complétée par celles obtenues par des approches probabilistes. Dans ce cadre, nous avons considéré le modèle Okapi (ou BM25) [ROB 00] utilisant la formulation suivante :

$$w_{ij} = [(k_1+1) \cdot tf_{ij}] / (\mathrm{K} + tf_{ij}) \qquad \text{avec } \mathrm{K} = k_1 \cdot [(1-b) + ((b \cdot l_i) / mean\ dl)] \qquad (1)$$

dans laquelle $l_i$ est la longueur du $i^{e}$ article (mesurée en nombre de termes d’indexation), et $b$, $k_1$ des constantes fixées empiriquement à $b = 0{,}55$, $k_1 = 1{,}2$ et *mean dl* la longueur moyenne d’un document.

Comme autres approches probabilistes, nous avons implémenté le modèle DFR-PL2 et le modèle DFR-I($n_e$)C2, issus de la famille *Divergence from Randomness* (DFR) [AMA 02]. Pour ces modèles, la pondération $w_{ij}$ combine deux mesures d'information, à savoir :

$$w_{ij} = \mathrm{Inf}^1_{ij} \cdot \mathrm{Inf}^2_{ij} = \mathrm{Inf}^1_{ij} \cdot (1 - \mathrm{Prob}^2_{ij}) \qquad (2)$$

Pour le modèle DFR-PL2, ces deux mesures d’informations sont estimées en recourant à la formulation suivante :

$$\mathrm{Prob}^2_{ij} = tfn_{ij} / (tfn_{ij} + 1) \quad \text{avec } tfn_{ij} = tf_{ij} \cdot \ln[1+((c \cdot mean\ dl) / l_i)]$$
$$\mathrm{Inf}^1_{ij} = -\log_2[(e^{-\lambda_j} \cdot \lambda_j^{tf_{ij}})/tf_{ij}!] \quad \text{avec } \lambda_j = tc_j / n \qquad (3)$$

dans laquelle $tc_j$ représente le nombre d'occurrences du $j^{e}$ terme dans la collection et $c$ une constante fixée empiriquement à 1,5.

Le modèle DFR-I($n_e$)C2 se base sur la formulation suivante.

$$Prob^2_{ij} = 1 - [(tc_j + 1) / (df_j \cdot (tfn_{ij}+1))]$$
$$Inf^1_{ij} = tfn_{ij} \cdot \log_2[(n+1) / (n_e+0{,}5)] \text{ avec } n_e = n \cdot [1 - [(n-1)/n]^{tc_j}] \quad (4)$$

Enfin, nous avons repris un modèle de langue (LM) [HIE 00], [HIE 02] dans lequel les probabilités sont estimées directement en se basant sur les fréquences d'occurrences dans le document D ou dans le corpus C dans son ensemble. Dans cet article, nous avons repris le modèle de Hiemstra [HIE 00] décrit dans l'équation 5 combinant une estimation basée sur le document (soit $Prob[t_j | D_i]$) et sur le corpus ($Prob[t_j | C]$).

$$Prob[D_i | Q] = Prob[D_i] \cdot \prod_{t_j \in Q} [\lambda_j \cdot Prob[t_j | D_i] + (1-\lambda_j) \cdot Prob[t_j | C]] \quad (5)$$
$$\text{avec } Prob[t_j | D_i] = tf_{ij} / nt_i \quad \text{et } Prob[t_j | C] = df_j / lc \quad \text{avec } lc = \sum_k df_k \quad (6)$$

dans laquelle $\lambda_j$ est un facteur de lissage (une constante pour tous les termes $t_j$, fixée à 0,35) et *lc* correspond à une estimation de la taille du corpus C.

Afin de comparer les résultats d'une analyse morphologique avec divers enracineurs, nous avons implémenté quatre algorithmes. Le premier correspond à un un enracineur léger visant à éliminer la marque du pluriel dans la langue anglaise (soit le suffixe '-s'). Cette solution comportant trois règles simples a été proposée par Harman [HAR 91] et sera noté "S-stemmer" dans la suite de cet article. D'autres enracineurs cherchent également à éliminer les suffixes dérivationnels (par exemple avec les suffixes '-ship', '-ability' ou '-ment') et nécessitent un nombre plus important de règles comme l'approche proposée par Lovins [LOV 68] (260 règles) et celle suggérée par Porter [POR 80] (60 règles). Le dernier enracineur testé a été repris du système de recherche d'information SMART [SAL 81]. Il correspond essentiellement à une version améliorée de l'algorithme de Lovins.

## 4. Evaluation

Pour mesurer la performance des différents modèles de recherche d'information, nous avons utilisé la précision moyenne (MAP ou *mean average precision*) obtenue par le système `trec_eval` [BUC 05]. Cette mesure a été adoptée par diverses campagnes d'évaluation pour mesurer la qualité de la réponse calculé par un système de dépistage de l'information. Afin de savoir si une différence entre deux modèles s'avère statistiquement significative, nous avons opté pour un test bilatéral non-paramétrique (basée sur le ré-échantillonnage aléatoire ou *bootstrap* [SAV 97], avec un seuil de signification $\alpha = 5$ %).

### *4.1 Evaluation des modèles de recherche et des enracineurs*

Se basant sur la méthodologie décrite ci-dessus, les résultats obtenus en utilisant quatre enracineurs et cinq modèles de dépistage sont décrits dans la table 1. Dans la deuxième colonne (dénommée "aucun") nous avons indiqué la performance obtenue en l'absence de tout traitement morphologique. Les quatre colonnes subséquentes représentent les quatre enracineurs choisis tandis que la dernière colonne (avec

l'étiquette “lemme”) montre les précisions moyennes que l'on obtient en utilisant une analyse morphologique. Toutes ces performances sont calculées avec des requêtes très courtes correspond aux titres des besoins d'information.

Dans les tables de cette étude, le meilleur résultat dans des conditions données est écrit en gras. En prenant cette valeur comme base pour notre test statistique, nous avons souligné dans chaque colonne les valeurs montrant une différence statistiquement significative. On remarque que, sauf pour l’enracineur léger de Harman (“S-stemmer”), la meilleure performance s'obtient toujours avec le modèle Okapi. Toutefois cette supériorité n'est pas statistiquement significative avec les modèles probabilistes DFR-PL2 et DFR-I($n_e$)C2 qui offrent des performances similaires. D'un autre côté, le modèle de langue (LM) et le modèle vectoriel fournissent des différences statistiquement significatives comparées à la précision moyenne du modèle Okapi.

| | Précision moyenne (MAP) | | | | | |
|---|---|---|---|---|---|---|
| | aucun | S-stemmer ‡ | Porter | Lovins | SMART | lemme |
| Okapi | **0,3743** | 0,4044† | **0,4150**†‡ | **0,3930** | **0.4152**†‡ | 0,3988† |
| DFR-PL2 | 0,3703 | 0,4006† | 0,4116†‡ | 0,3927† | 0,4096†‡ | **0,3994**† |
| DFR-I($n_e$)C2 | 0,3731 | **0,4054**† | 0,4141†‡ | 0,3894 ‡ | 0,4139†‡ | 0,3988† |
| LM | 0,3445 | 0,3709† | 0,3809†‡ | 0,3522 ‡ | 0,3760†‡ | 0,3602 |
| *tf idf* | 0,2230 | 0,2393† | 0,2399† | 0,2194 ‡ | 0,2431†‡ | 0,2308 |
| Moyenne | 0,3370 | 0,3641 | 0,3723 | 0,3493 | 0,3716 | 0,3576 |
| % différence | | +8,0 % | +10,5 % | +3,6 % | +10,2 % | +6,1 % |

**Table 1. Précision moyenne (MAP) obtenues avec différents modèles et enracineurs (284 requêtes, format T)**

Si on prend comme base la précision moyenne obtenue sans aucun traitement lexical (colonne “aucun”), nous avons indiqué avec le symbole “†” les différences statistiquement significatives. Nous pouvons en conclure que dans presque tous les cas, le recours à un enracineur améliore significativement les performances.

Dans un troisième temps, on souhaite connaître s'il existe des différences significatives entre les performances de divers enracineurs ou par rapport à une analyse morphologique. Dans ce but, nous sélectionnons comme valeurs de référence la performance de l'enracineur léger (“S-stemmer”) et toute différence significative sera signalée par le symbole “‡”. Dans ce cas, on observe que l'algorithme proposé par Porter ou celui utilisé par SMART fournissent une meilleure performance. Par contre, l’approche plus agressive suggérée par Lovins donne, significativement, de moins bons résultats en moyenne. Enfin, les différences de

performance entre l'enracineur léger et l'analyse morphologique ne s'avèrent pas statistiquement significatives.

A titre de comparaison, Harman [HAR 91] indique que, basé sur le modèle *tf idf*, aucune différence de performance statistiquement significative n'a pu être détectée entre les enracineurs proposés par Porter [POR 80], Lovins [LOV 68] ou le S-stemmer [HAR 91]. Pour Hull [HUL 96] s'appuyant sur une variante du modèle classique *tf idf*, une amélioration faible (de l'ordre de 1 % à 3 %) peut être attribuable à l'emploi d'enracineurs. Selon cette étude, tous les enracineurs proposent une précision moyenne significativement supérieure à un modèle renonçant à cette forme de normalisation lexicale. De plus, le S-stemmer offrirait une qualité inferieure à celle obtenue par les algorithmes de Porter [POR 80] ou de Lovins [LOV 68]. Ces deux études se limitaient au modèle classique *tf idf* et plus, elles se basaient sur des collections différentes et disposaient d'un nombre de requêtes plus restreint.

### *4.2 Analyse morphologique, partie du discours et thésaurus*

Un des avantages indéniables de la campagne CLEF 2008 consistait à mesurer l'impact de l'analyse morphologique mais également à vérifier l'influence de deux informations pouvant, du moins en partie, désambiguïser la sémantique attachée aux mots. En effet, nous pouvons utiliser les diverses étiquettes concernant la partie du discours (POS) et celles associées au numéro de la classe du thésaurus (synset).

Afin de disposer d'une base de comparaison plus large, nous avons évalué nos approches avec des requêtes courtes (T) dans la table 2 d'une part et, d'autre part, avec des requêtes de taille moyenne (TD), valeurs de performances reportées dans la table 3. Dans les deux cas, la deuxième colonne indique la précision moyenne (MAP) obtenue en utilisant le lemme lors de l'indexation des documents et des requêtes (même valeur que dans la table 1, dernière colonne).

Dans la troisième colonne, le score des documents dépistés sera augmenté si le lemme commun entre eux et la requête possède la même partie du discours (POS). Cette information peut déterminer plus précisément le sens attaché à un terme. En anglais par exemple, le même mot peut avoir des sens différents, comme par exemple "lean" (maigre) comme adjectif ou comme verbe (incliner). Le mot "face" (ou "form", "mean") possède aussi un sens quelque peu distinct lorsque on l'utilise en tant que nom (visage) ou que verbe (assumer, faire face). Dans le but de tenir compte de cette information, nous rajoutons pour chaque terme d'indexation une deuxième unité d'indexation composée du lemme et de son étiquette POS (dérivé du projet *Penn Treebank* [MAR 93]). Par exemple pour l'adjectif "white" on collerait son POS (JJ) et obtiendrait l'unité d'indexation "whiteJJ". Si par contre on rencontre le nom propre "White" ou le nom commun on obtiendrait l'unité d'indexation "whiteNNP" respectivement "whiteNN" où NNP indique qu'il s'agit un nom propre et NN d'un nom commun. Ces deux unités d'indexation diffèrent de "whiteJJ" et permettent alors de distinguer les trois utilisations différentes de la racine "white".

Dans la dernière colonne nous avons reporté la performance obtenue lorsque l'on accroit le score d'un document extrait si le terme commun entre celui-ci et la requête possède le même numéro synset. Dans cette perspective, tous les numéros de classe du thésaurus sont ajoutés à la représentation des documents et des requêtes.

Comme dans la table 1, les valeurs imprimées en gras signalent la meilleure performance pour une colonne donnée, et les performances soulignées indiquent les différences statistiquement significatives par rapport au meilleur score d'une colonne. Les modèles probabilistes DFR ou Okapi proposent une meilleure qualité qui s'avère statistiquement différente de celle obtenue par le modèle de langue (LM) ou l'approche vectorielle.

Si l'on adopte comme référence les performances de la deuxième colonne, les différences statistiquement significatives seront signalées par le symbole "†" placé après la valeur analysée. Dans la table 2 (requêtes T), le modèle Okapi ou LM propose une variation significative de la performance moyenne lorsque l'on tient compte de la partie du discours (POS).

| | Précision moyenne (MAP) | | |
|---|---|---|---|
| | lemme | lemme & POS | lemme & synset |
| Okapi | 0,3988 | **0,4053**† | 0.3986 |
| DFR-PL2 | **0,3994** | 0,4013 | 0,3918 |
| DFR-I($n_e$)C2 | 0,3988 | 0,4047 | **0,4018** |
| LM | 0,3602 | 0,3659† | 0,3546 |
| *tf idf* | 0,2308 | 0,2315 | 0,2325 |
| Moyenne | 0,3576 | 0,3617 | 0,3559 |
| % différence | | +1,2 % | -0,5 % |

**Table 2. Précision moyenne (MAP) pour différents modèles de RI et variantes d'analyse morphologique (284 requêtes, format T)**

Si l'on analyse quelques requêtes, nous pouvons mieux comprendre l'effet de cette information lors de la recherche. Avec le modèle Okapi par exemple, on observe une amélioration de la moyenne de 0,3988 à 0,4053 lors de la prise en compte des parties du discours. Dans ce cas, on améliore la performance pour 133 requêtes, mais on constate une détérioration pour 108 interrogations (il n'y a pas de changement pour le solde des 44 requêtes). L'interrogation n° 217 ("AIDS in Africa") propose la plus grande variation. Dans ce cas, la précision moyenne passe de 0,2037 lorsque l'on ignore les étiquettes POS à 0,5556. Lors du traitement de la requête, le système convertit "AIDS" en "aid" augmentant ainsi le nombre de correspondances possibles ("aid" possédant d'autre sens dans le corpus). Avec la

prise en compte de la partie du discours, “aid” est signalé comme nom propre (étiquette NNP), et ainsi les documents contenant cette abréviation verront leur similarité avec la requête s'accroitre et leur classement s'améliorer.

| | Précision moyenne (MAP) | | |
|---|---|---|---|
| | lemme | lemme & POS | lemme & synset |
| Okapi | 0,4663 | 0,4720† | 0,4395† |
| DFR-PL2 | 0,4608 | 0,4634 | 0,4365† |
| DFR-I($n_e$)C2 | **0,4671** | **0,4740**† | **0,4665** |
| LM | 0,4444 | 0,4562† | 0,4342† |
| *tf idf* | 0,2778 | 0,2879† | 0,2834 |
| Moyenne | 0,4597 | 0,4664 | 0,4442 |
| % différence | | +1,5 % | -3,4 % |

**Table 3. Précision moyenne (MAP) pour différents modèles de RI et variantes d’analyse morphologique (284 requêtes, format TD)**

Les évaluations reportées dans la table 3 ont été obtenues avec des requêtes de taille moyenne (TD). Dans ce contexte, le modèle DFR-I($n_e$)C2 fournit la meilleure performance (précision moyenne imprimée en gras). Comme dans les évaluations précédentes, les différences avec le modèle de langue (LM) ou l'approche vectorielle *tf idf* sont statistiquement significatives (valeurs soulignées). Si les valeurs moyennes obtenus par l’analyse morphologique (colonne “lemme”) servent de référence, on observe que, exception faite du modèle DFR-PL2, toutes les autres modèles recourant aux étiquettes POS apportent des performances moyennes statistiquement significatives (ajout du symbole “†”), même si ces différences demeures faibles en valeur absolue.

Si on analyse quelques cas, nous constatons que pour le modèle DFR-I($n_e$)C2 l'emploi des informations POS permet d'accroitre la MAP de 0,4671 à 0,4740. Derrière cette variation, on observe une amélioration pour 138 interrogations et une dégradation pour 98 requêtes. Comme pour les requêtes courtes, l'interrogation n$^o$ 217 (“AIDS in Africa”) voit sa précision moyenne passer de 0,1944 (“lemme”) à 0,5526 (“lemme & POS”).

L'ajout des numéros de classe du thésaurus (“lemme & synset”) apporte des dégradations de performance qui s'avèrent significatives pour les modèles Okapi, DFR-PL2 et LM. Pour expliquer ce phénomène, nous pouvons analyser quelques requêtes. Par exemple pour le modèle Okapi et l'interrogation n$^o$ 76 (“Solar Energy”), la précision moyenne passe de 0,663 (“lemme”) à 0,0722 avec la prise en compte des numéros de classe du thésaurus. Pour cette requête, sa partie descriptive contient la forme “is” et deux occurrences de “being”. Le lemme correspondant “be” engendre dix numéros de synsets qui s'ajoutent à la représentation interne. Ainsi chaque document contenant une forme verbale du verbe “to be” aura à chaque

fois dix appariements avec la requête rendant plus difficile la discrimination entre les documents pertinents et ceux qui en le sont pas. *A posteriori* on pourrait imaginer utiliser une liste de mots outils afin d'éliminer les termes ayant une haute fréquence dans la collection avec leur synset pour éviter ce genre de problème. Or on doit aussi noter que ceci aurait probablement un grand effet sur un nombre très restreint de requêtes, mais sur l'ensemble de requêtes l'amélioration serait probablement négligeable.

## 5. Conclusion

Sur la base d'un corpus rédigé en langue anglaise et enrichi d'information morphologique, nous avons démontré que les meilleures performances s'obtiennent avec le modèle Okapi ou DFR-I($n_e$)C2. Cependant, les différences de précision moyenne entre ces deux modèles ou la variante DFR-PL2 ne sont pas significatives. D'un autre côté, le modèle de langue (LM) et le modèle vectoriel *tf idf* produisent des différences statistiquement significatives comparées à la précision moyenne la plus élevée.

Quelque soit le modèle de recherche considéré, l'emploi d'un enracineur ou d'un traitement morphologique permet d'améliorer significativement la performance moyenne. Entre ces diverses approches, nous avons observé une différence significative entre l'algorithme de Porter ou celui du système SMART et un enracineur léger noté S-stemmer. Avec l'approche suggérée par Lovins, la performance moyenne se détériore comparée aux autres enracineurs. L'indexation par lemmes donne des résultats similaires aux algorithmes de Porter ou du système SMART. On constate en particulier que l'approche très simple de l'enracineur léger basé sur trois règles ayant comme but d'éliminer les marques du pluriel (S-stemmer), s'avère au moins aussi efficace que des approches plus agressives comme celles proposés par Lovins (260 règles) et par Porter (60 règles).

Pour favoriser la transparence de ce processus envers l'utilisateur, il serait avantageux d'utiliser une approche simple, produisant des performances similaires à des approches plus agressives. Ces dernières tendent à réduire plus fortement les formes de surface, rendant parfois moins compréhensible l'appariement entre formes de surface et racine pour l'utilisateur. Dans cette optique, le moteur de recherche Google applique un enracineur léger. Ainsi si l'on soumet la requête "computes" le système retourne des documents contenant aussi les formes "compute", "computers" ou "computing" mais pas la forme "computable".

L'adjonction de la partie du discours permet, dans le cadre de requête moyenne (TD), d'augmenter significativement la performance moyenne. Avec des interrogations de faible longueur (T), cette amélioration s'avère significative pour le modèle de langue et l'approche Okapi. L'inclusion des numéros de classe d'un thésaurus dans les documents et les requêtes a tendance à diminuer la précision

moyenne. En présence de requêtes de taille moyenne (TD), la différence de performance s'avère souvent significative. Notre solution doit toutefois être vue comme un premier essai qui mériterait quelques améliorations comme la sélection d'une seule classe par mot au lieu de laisser l'ensemble des possibilités. Comme autre possibilité, nous pourrions tenir compte du thésaurus uniquement pour certaines parties du discours comme les noms par exemple.

En plus il faut remarquer que les résultats obtenus sont liés à la langue anglaise, et pourrait s'appliquer à d'autres langues possédant une morphologie flexionnelle simple. Tomlinson [TOM 04] par exemple compare les enracineurs lexicaux et algorithmiques pour neuf langues européennes (allemand, français, italien, espagnol, néerlandais, finnois, suédois, russe et anglais). Pour le finnois et l'allemand, l'analyse morphologique (enracineur lexical) apporte des performances significativement supérieures tandis que pour les autres langues, les différences de performances ne sont pas significatives. La présence d'une morphologie flexionnelle complexe semble être à l'origine de ces différences.

Remerciements

Cette recherche a été financée en partie par le Fonds national suisse pour la recherche scientifique (subside n$^0$ 200021-113273).

## 6. 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$^o$ 4, 2002, p. 357-389.

[BUC 05] Buckley, C., & Voorhees, E.M. "Retrieval system evaluation", E.M. Voorhees, D.K. Harman (Eds): *TREC. Experiment and Evaluation in Information Retrieval* (pp. 53-75). The MIT Press, Cambridge (MA), 2005.

[CHA 07] Chan, Y.S., Ng, H.T., & Zhong, Z. "NUS-PT: Exploiting Parallel Texts for Word Sense Disambiguation in the English All-Words Tasks", *Proceedings of the 4th International Workshop on Semantic Evaluations* (SemEval 2007), Prague, p. 253–256.

[FEL 98] Fellbaum, C. "*WordNet. An Electronic Lexical Database*", The MIT Press, Cambridge (MA), 1998.

[FOX 90] Fox, C. "A stop list for general text", *ACM-SIGIR Forum*, vol. 24, n$^o$ 1, 1990, p. 19-35.

[HAR 91] Harman, D. "How effective is suffixing?", *Journal of the American Society for Information Science*, vol. 42, n$^o$ 1, 1991, p. 7-15.

[HIE 00] Hiemstra, D. "Using language models for information retrieval. CTIT Ph.D. Thesis, 2000.

[HIE 02] Hiemstra, D. "Term-specific smoothing for the language modeling approach to information retrieval. The importance of a query term", *Proceedings of the ACM-SIGIR'2002*, Tempere, p. 35-41.

[HUL 96] Hull, D. “Stemming algorithms: A case study for detailed evaluation”, *Journal of the American Society for Information Science*, vol. 47, n° 1, 1996, p. 70-84.

[LOV 68] Lovins, J.B. “Development of a stemming algorithm”, Mechanical Translation and Computational Linguistics, vol. 11, n° 1, 1968, p. 22-31

[MAN 08] Manning, C.D, Raghavan, P., & Schütze, H. “*Introduction to Information Retrieval*”, Cambridge University Press, Cambridge (UK), 2008.

[PER 08] Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A. & Santos, D. (Eds.). “*Advances in Multilingual and Multimodal Information Retrieval*”, LNCS #5152, Springer-Verlag, Berlin, 2008.

[POR 80] Porter, M.F. “An algorithm for suffix stripping”, *Program*, vol. 14, n° 3, 1980, p. 130-137.

[RAT 96] Ratnaparkhi, A. “A maximum entropy part-of-speech tagger”, *Proceedings of the Empirical Methods in Natural Language Processing Conference*, 1996, p. 133-142.

[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.

[SAL 71] Salton, G. “*The SMART Retrieval System - Experiments in Automatic Document Processing*”, Prentice-Hall Inc., Englewood Cliffs (NJ), 1971.

[SAV 93] Savoy, J. “Stemming of French words based on grammatical category”, *Journal of the American Society for Information Science*, vol. 44, n° 1, 1993, p. 1-9.

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

[TOM 04] Tomlinson, S. “Lexical and algorithmic stemming compared for 9 European languages with Hummingbird SearchServer™ at CLEF 2003”, *Comparative Evaluation of Multilingual Information Access Systems* LNCS #3237, Springer-Verlag, Berlin, 2004. p. 286-300.