Extension de requêtes par relations morphologiques acquises automatiquement

Fabienne Moreau, Vincent Claveau

Extension de requêtes par relations morphologiques acquises automatiquement

Fabienne Moreau, Vincent Claveau

IRISA - CNRS
Campus universitaire de Beaulieu
Avenue du Général Leclerc
35042 Rennes cedex, France

{Fabienne.Moreau,Vincent.Claveau@irisa.fr}

Résumé

Cet article s'intéresse à la prise en compte de la variation morphologique en recherche d'information. L'approche proposée est une méthode simple de reconnaissance des variantes morphologiques utilisées pour l'enrichissement des requêtes au sein d'un système de recherche d'information (SRI). À l'inverse de nombreux travaux existants, la technique proposée présente la particularité de ne nécessiter aucunes ressources ni connaissances externes, et d'être ainsi applicable à une grande variété de langues. Les évaluations de cette approche réalisées sur plusieurs collections de documents, sur 6 langues et comparées à différents outils existants (stemmer, lemmatiseur) attestent de l'intérêt de la méthode puisqu'une amélioration significative des performances des SRI est constatée dans tous les cas.

Mots-clés *: Variation morphologique, extension de requêtes, apprentissage automatique par analogie.*

Abstract

Information retrieval systems (IRS) usually suffer from a low ability to recognize a same idea or concept that is expressed in different forms. A way of improving these systems is to take into account morphological variations. In this paper, we propose a simple method to recognize these variations that are further used so as to enrich queries. In comparison with already published methods, this system does not need an external ressources or knowledge and thus supports many languages. This approach is evaluated on several collections, 6 different languages and compared to existing tools (stemmer, lemmatizer) ; reported results show a significant improvement of the overall IRS performance in every case.

Key-words*: Morphological variation, query expansion, analogy-based machine learning.*

1 INTRODUCTION

La principale difficulté des systèmes de recherche d'information (SRI) est de faire correspondre l'information recherchée par l'utilisateur (par le biais d'une requête en langage naturel) avec celle contenue dans les documents. La méthode généralement utilisée repose sur une mise en correspondance des mots utilisés dans la requête avec ceux représentant le contenu des documents. À partir de ce mécanisme d'appariement fondé sur une simple comparaison de chaînes de caractères, les SRI se trouvent rapidement confrontés à deux problèmes liés à la complexité du langage naturel. Le premier est lié à la polysémie : un même terme peut référer à des concepts différents. Cette ambiguïté des termes entraîne potentiellement la récupération par le SRI de documents non pertinents. La seconde difficulté, duale de la première, est liée à la possibilité de formuler de différentes manières un même concept. Un document pertinent peut ainsi contenir des termes « sémantiquement proches » de ceux de la requête mais toutefois différents (synonymes, hyperonymes...). Ce phénomène empêche les systèmes de retourner à l'utilisateur des documents pourtant pertinents.

Cet article présente une approche pour contourner en partie ce second problème par la prise en compte de la variation morphologique. En effet, une relation morphologique entre deux mots est souvent l'indice d'une relation sémantique. La connaissance de ces relations permettrait ainsi de manipuler des énoncés sémantiquement proches bien que différents en apparence. Par exemple, cela rendrait possible l'appariement d'une requête contenant le terme *déménager* et d'un document contenant le mot *déménagement*. L'objectif de cette prise en compte des phénomènes morphologiques est donc au final d'améliorer les performances (rappel, précision...) des SRI.

Le problème de la variation morphologique est bien connu en RI et de nombreux travaux y ont été consacrés. Parmi ceux-ci, beaucoup ont été développés pour une langue donnée et s'appuient sur des connaissances externes (règles de réécriture, bases de suffixes, lexiques...), ce qui limite leur réutilisabilité hors de leur cadre de développement. Dans cet article, pour répondre en partie à ce problème de portabilité, nous présentons et évaluons une approche simple et efficace pour le traitement de certaines variantes morphologiques en RI avec les hypothèses suivantes :

- le système ne doit nécessiter aucunes connaissances ou données externes ;
- le système doit être complètement automatique ;
- le système doit pouvoir s'appliquer directement à diverses langues.

En accord avec ces hypothèses, la technique proposée repère automatiquement au sein de la collection de textes traitée par le SRI des mots en relations morphologiques avec les termes de la requête et les ajoute à cette dernière.

La section suivante présente brièvement la variation morphologique et quelques uns des travaux existants pour gérer ce type de variation en RI. Nous détaillons ensuite le fonctionnement de notre approche et son utilisa-

tion pour l'extension de requêtes en section 3. Nous en évaluons les résultats sur des collections de textes et requêtes dans la section 4 avant de conclure et de proposer quelques perspectives ouvertes par ces travaux.

2 VARIATION MORPHOLOGIQUE EN RI

La variation morphologique a fait l'objet de nombreux travaux en RI. Un état de l'art très complet peut être trouvé dans [24]. Nous n'en donnons dans cette section que les grandes lignes et le principe et nous concentrons sur les travaux les plus proches des nôtres.

2.1 Le phénomène de la variation morphologique

Dans de nombreuses langues, certains mots partagent une proximité graphique et sémantique ; on parle de relations morphologiques. Ainsi en français, le verbe *transformer* est ainsi lié à *transformes, transforme, transformateur, transformation*... On distingue usuellement plusieurs types de relations morphologiques [23] ; celles qui nous intéressent dans cet article sont la flexion et la dérivation. La flexion est la relation existant entre deux mots que seuls distinguent le nombre, le genre, le temps, personne et mode pour les verbes, ou le cas pour les langues à déclinaisons (allemand, polonais, russe...). Les différents mots liés par la flexion sont appelés formes fléchies ; parmi celles-ci on choisit souvent un unique représentant, le lemme. Par exemple, pour les verbes en français, le lemme est la forme infinitive, pour les adjectifs, le masculin singulier... Une autre relation morphologique souvent manipulée en RI est la dérivation. En morphologie dérivationnelle, deux mots reliés morphologiquement possèdent une racine commune, et diffèrent par leurs affixes (principalement des préfixes comme *re-* dans *reconstruire* ou des suffixes comme *-eur* dans *constructeur*, mais aussi des infixes dans certaines langues). Cette dérivation s'accompagne alors d'une modification légère de sens (comme dans *faire* $\leftrightarrow$ *défaire*) et/ou de catégories grammaticales (*décider* $\leftrightarrow$ *décision*).

Des termes qui dérivent du même lemme ou de la même racine présupposent donc généralement un sens proche. En ce sens, la variation morphologique constitue un type particulier de variation sémantique qu'il est intéressant de capturer, notamment en RI. La prise en compte de la variation morphologique a fait l'objet de nombreuses expérimentations dans ce domaine. Nous proposons ci-dessous un rappel des principales méthodes utilisées et des expériences au sein desquelles elles sont appliquées.

2.2 Les outils morphologiques en RI

Au sein d'un SRI, les variantes morphologiques sont prises en compte soit lors de la phase d'indexation des documents et des requêtes (approche par

conflation), soit pour l'enrichissement de ces dernières (approche par extension ou expansion de requêtes). Pour la conflation, les différentes variantes morphologiques possibles d'un mot sont normalisées, *i.e.* ramenées à une seule et même forme, racine ou lemme, lors de l'indexation. L'appariement des documents et de la requête se fait alors sur la base de cette forme canonique. Pour l'extension de requêtes, il s'agit de procéder également à la reconnaissance des variantes morphologiques sans opérer cependant de traitement de normalisation. Les termes de la requête de l'utilisateur sont enrichis par le biais de leurs variantes morphologiques au moment de la recherche.

Une approche très commune, le *stemming* (ou racinisation), cherche à rassembler les différentes variantes d'un mot autour d'un *stem* (*i.e.* une pseudo-racine). Cette procédure traite à la fois des cas relevant de la flexion et de la dérivation. Les techniques utilisées pour procéder à la racinisation reposent généralement sur une liste d'affixes de la langue considérée et sur un ensemble de règles de désuffixation construites *a priori* [28, 21] qui permettent, étant donné un mot de trouver son *stem*. Bien que le principal avantage de ces outils réside dans leur simplicité, l'absence de contraintes linguistiques fortes engendre néanmoins des erreurs de sur-racinisation (*e.g.* le *stem nat* qui regroupe à la fois *nature* et *nation*) ou de sous-racinisation (*e.g.* le *stem adaptat* qui empêche le regroupement des formes *adapter* et *adaptation*). L'impact du *stemming* sur les performances des SRI est cependant mitigé. Les expériences sur l'anglais sont plutôt décevantes [19, 12, 7] et variables selon les collections [17] et la taille des requêtes [15]. L'enrichissement des requêtes par des variantes morphologiques issues de *stemming* a également abouti à des améliorations assez faibles sur l'anglais [5].

D'autres expériences en RI ont tenté de prendre en compte la variation morphologique à l'aide d'outils plus « sophistiqués » que les *stemmers*, s'appuyant notamment sur des outils ou des données issus du traitement automatique des langues. Des outils performants de lemmatisation existent maintenant pour de nombreuses langues. Les outils d'analyses dérivationnelles, bien que moins courants et parfois attachés à un domaine (par exemple le domaine biomédical), sont aussi parfois disponibles. En RI, ces outils sont le plus souvent utilisés successivement, la lemmatisation précédant l'analyse dérivationnelle, et apportent un gain de performances variable selon les langues et les expériences mais globalement positif [27, 33, 32, entre autres].

Ainsi, bien que la prise en compte des variantes morphologiques en RI apparaît être évidente pour l'amélioration des performances des SRI, le choix de la méthode la plus adaptée pour le traitement des variantes morphologiques en RI reste particulièrement difficile. Un certain nombre de travaux ont cherché à comparer l'efficacité des raciniseurs par rapport aux analyseurs fondés sur des traitements linguistiques [16, 25]. Néanmoins, les conclusions sont confuses et il reste difficile d'affirmer la supériorité de tel outil par rapport à tel autre.

2.3 Travaux connexes

Parmi les travaux évoqués précédemment, peu sont compatibles avec les trois hypothèses données en introduction comme cadre de nos travaux. En effet, la plupart des outils utilisés pour gérer la variation morphologique (*stemmer*, lemmatiseur...) reposent sur des connaissances externes, propres à la langue traitée ; mais quelques travaux existants s'inscrivent cependant dans notre cadre.

Quelques auteurs ont ainsi développé des approches devant permettre de capturer les phénomènes morphologiques à l'œuvre dans diverses langues, et ce de manière non-supervisée [10, 18], le plus souvent à l'aide d'approches statistiques. Certaines de ces techniques ont ensuite été appliquées à la recherche d'information mais ne semblent apporter que des gains limités [11, 26]. Une autre approche est adoptée par Xu et Croft [35] : ils s'appuient sur les cooccurrences collectées sur la collection de textes de leur SRI pour améliorer, mais seulement de quelques pourcents, les performances de *stemmers* basiques. Les systèmes d'indexation par n-gram [6] peuvent également être vus comme des techniques permettant de contourner le problème de la variation morphologique de manière non-supervisée. Malheureusement, cette approche un peu brutale n'apporte en pratique pas ou peu d'amélioration des résultats [32].

Un travail très proche du nôtre est celui de Gaussier et collègues [8, 9] concernant l'utilisation d'une technique de désuffixation en RI. Par le biais d'une méthode d'apprentissage non supervisée des suffixes et des opérations de suffixation à partir de lexiques flexionnels de la langue, le système proposé permet l'extraction de suffixes potentiels ensuite utilisés pour relier les lemmes de même famille. Plusieurs différences existent entre nos travaux et ceux-ci. D'une part, en plus des suffixes, notre approche prend en compte les préfixes, ce qui permet la découverte de certaines relations morphologiques intéressantes (*e.g.* *foudre* $\leftrightarrow$ *parafoudre*, *cancéreux* $\leftrightarrow$ *anticancéreux* ; voir l'évaluation de l'intérêt de cette préfixation pour notre tâche de RI en section 4.2). D'autre part, les travaux de Gaussier et collègues tirent parti d'informations catégorielles (nom, verbe, adjectif...) sur les mots manipulés. Une telle liste implique soit l'emploi de données externes, soit d'outils dédiés, ce qui est incompatible avec nos hypothèses ; notre approche n'utilise que les textes manipulés par le SRI sans aucun pré-traitement. Enfin, lors de son application à la RI, le système proposé par Gaussier et collègues fonctionne par conflation : les mots détectés comme étant liés morphologiquement sont réunis en une seule forme qui est utilisée pour l'indexation. Pour notre part, nous utilisons les variantes morphologiques uniquement pour étendre les requêtes ; ce mode d'utilisation est plus souple et mieux adapté à notre système (voir la discussion en section 4.5 sur ce sujet), rejoignant en cela les conclusions de certains auteurs [35, 2].

3 ACQUISITION DES VARIANTES MORPHOLOGIQUES EN RI

Dans cette section, nous présentons dans un premier temps la technique d'acquisition de variantes morphologiques utilisée. Dans un deuxième temps, nous détaillons son utilisation effective au sein d'un SRI pour étendre les requêtes avec les variantes morphologiques trouvées.

3.1 Acquisition par analogies

L'approche que nous avons adoptée pour acquérir les variantes morphologiques des mots contenus dans les requêtes s'appuie sur une technique développée initialement à des fins terminologiques [4, 3]. Le principe de cette technique d'acquisition morphologique est relativement simple et s'appuie sur la construction d'analogies. En toute généralité, une analogie peut être représentée formellement par la proposition A : B $\doteq$ C : D, qui signifie « A est à B ce que C est à D » ; c'est-à-dire que le couple A-B est en analogie avec le couple C-D. Son utilisation en morphologie, assez évidente, a déjà fait l'objet de plusieurs travaux [13, 20] : par exemple, si l'on postule l'analogie

connecteur : *connecter* $\doteq$ *éditeur* : *éditer*

et si l'on sait par ailleurs que *connecteur* et *connecter* partagent un lien morpho-sémantique, on peut alors supposer qu'il en est de même pour *éditeur* et *éditer*.

Le préalable essentiel à l'utilisation effective de l'apprentissage par analogie est la définition de la notion de similarité qui permet de statuer que deux paires de propositions – dans notre cas deux couples de mots – sont en analogie. La notion de similarité que nous utilisons, notée Sim, est simple mais adaptée aux nombreuses autres langues dans lesquelles la flexion et la dérivation sont principalement obtenues par préfixation et suffixation.

Intuitivement, Sim vérifie que, pour passer d'un mot m_3 à un mot m_4, les mêmes opérations de préfixation et de suffixation que pour passer de m_1 à m_2 sont nécessaires. Plus formellement, notons $\mathrm{lcss}(\mathrm{X}, \mathrm{Y})$ la plus longue sous-chaîne commune à deux chaînes de caractères X et Y (*e.g.* lcss(*installer*, *désinstallation*) = *install*), et $\mathrm{X} +_{suf} \mathrm{Y}$ (respectivement $+_{pre}$) la concaténation du suffixe (resp., préfixe) Y à X, et $\mathrm{X} -_{suf} \mathrm{Y}$ (respectivement $-_{pre}$) la soustraction du suffixe (resp., préfixe) Y à X. La mesure de similarité Sim est alors définie de la manière suivante :

Sim(m_1-m_2, m_3-m_4) = 1 si on a simultanément les quatre conditions :

$$\begin{cases} m_1 = \mathrm{lcss}(m_1, m_2) +_{pre} \mathrm{Pre}_1 +_{suf} \mathrm{Suf}_1, \text{ et} \\ m_2 = \mathrm{lcss}(m_1, m_2) +_{pre} \mathrm{Pre}_2 +_{suf} \mathrm{Suf}_2, \text{ et} \\ m_3 = \mathrm{lcss}(m_3, m_4) +_{pre} \mathrm{Pre}_1 +_{suf} \mathrm{Suf}_1, \text{ et} \\ m_4 = \mathrm{lcss}(m_3, m_4) +_{pre} \mathrm{Pre}_2 +_{suf} \mathrm{Suf}_2 \end{cases}$$

Sim(m_1-m_2, m_3-m_4) = 0 sinon

où $\mathrm{Pre_i}$ et $\mathrm{Suf_i}$ sont des chaînes de caractères quelconques. Si Sim(m_1-m_2, m_3-m_4) = 1, cela signifie que l'analogie $m_1 : m_2 \doteq m_3 : m_4$ est vérifiée et donc on suppose que la relation morpho-sémantique entre m_1 et m_2 est la même qu'entre m_3 et m_4.

Notre processus de détection de variantes morphologiques consiste ainsi à vérifier, au moyen de la mesure Sim, si un couple de mots inconnus est en analogie avec un ou plusieurs exemples de couples connus. Par exemple, nous pouvons déterminer que le couple *déshydrater - réhydratation* (m_3-m_4) est en analogie avec le couple-exemple *désinstaller - réinstallation* (m_1,m_2), puisque, sachant que lcss(*désinstaller*, *réinstallation*) = *install* et que lcss(*déshydrater*, *réhydratation*) = *hydrat*, on a bien :

$$\begin{cases} m_1 = \textit{install} +_{pre} \textit{dés} +_{suf} \textit{er}, \text{ et} \\ m_2 = \textit{install} +_{pre} \textit{ré} +_{suf} \textit{ation}, \text{ et} \\ m_3 = \textit{hydrat} +_{pre} \textit{dés} +_{suf} \textit{er}, \text{ et} \\ m_4 = \textit{hydrat} +_{pre} \textit{ré} +_{suf} \textit{ation} \end{cases}$$

On peut donc noter : Sim(*désinstaller - réinstallation* , *déshydrater - réhydratation*) = 1.

En pratique, pour des raisons d'efficacité lors de la recherche d'analogies, plutôt que les couples-exemples, ce sont les opérations de préfixation et suffixation à l'œuvre dans la mesure de similarité Sim qui sont stockées. Ainsi, le couple-exemple *désinstaller* ↔ *réinstallation* n'est pas stocké en tant que tel, mais on conserve la règle :

$$m_2 = m_1 -_{pre} \textit{dés} +_{pre} \textit{ré} -_{suf} \textit{er} +_{suf} \textit{ation}$$

Montrer l'analogie *déshydrater* : *réhydratation* $\doteq$ *désinstaller* : *réinstallation* revient alors simplement à tester que *déshydrater* ↔ *réhydratation* vérifie la règle précédente.

Comme le soulignait É. Gaussier pour ses travaux [8], les opérations de suffixation et de préfixation en œuvre dans notre approche permettent de gérer en partie les variations légères de racines, pourvu qu'elles soient assez communes pour être présentes dans un de nos exemples. Ainsi, si l'on a dans notre base d'exemples le couple *récupère* ↔ *récupération*, le couple *agglomère* ↔ *agglomération* sera facilement reconnu, malgré le changement de racine *agglomér-/agglomèr-*, puisque l'on a bien l'analogie *récupère* : *récupération* $\doteq$ *agglomère* : *agglomération*. Bien entendu, les variations plus importantes comme celle existant dans le couple *foie* ↔ *hépatique* sont hors de portée de notre approche.

Nous avons montré que cette simple approche par analogie permettait ainsi de trouver des dérivés morphologiques à l'aide d'exemples de mots en relation morpho-sémantique avec une très bonne couverture et une grande précision dans un contexte de construction de terminologies (voir [3] pour le détail des résultats). Nous avons aussi montré qu'il était possible d'identifier avec d'excellents taux de réussite le lien sémantique précis entre ces dérivés en associant à chaque règle une ou plusieurs étiquettes de relation sémantique. Ce dernier point ne sera pas utilisé ici ; nous faisons l'hypothèse

que tous les liens sémantiques (synonymie, antonymie, hyperonymie...) sont pertinents pour notre application à l'extension de requêtes.

3.2 Utilisation dans le SRI

La technique de détection de dérivés morphologiques par analogie présentée ci-avant requiert des exemples de couples de mots morphologiquement liés pour pouvoir fonctionner. Cet aspect supervisé n'est pas adapté à une utilisation en RI et à nos hypothèses exposées en introduction ; on souhaite au contraire une totale autonomie du système. Pour répondre à ce problème, nous remplaçons cette phase de supervision humaine par une technique rustique permettant de constituer automatiquement un ensemble de paires de mots pouvant servir d'exemples.

Cette première phase de recherche de couples-exemples se déroule de la façon suivante :

1 – choisir un article au hasard dans la collection de textes du SRI ;

2 – constituer tous les couples de mots possibles (issus de l'article) ;

3 – ajouter aux exemples les couples m_1-m_2 tels que lcss(m_1,m_2) $> l$;

4 – retourner en 1.

On répète ces étapes jusqu'à obtenir un ensemble de couples-exemples jugé suffisamment important. Dans les expériences présentées en section 4, ce sont 500 articles qui ont été analysés ainsi.

Cette phase de constitution d'exemples repose donc sur la même hypothèse que précédemment : la dérivation et la flexion se font principalement par des opérations de préfixation et suffixation. Il n'est pas grave lors de cette phase de ne pas repérer des couples de mots morphologiquement liés ; cependant, pour le bon fonctionnement des analogies qui vont en être tirées, il faut éviter de constituer des couples qui ne seraient pas des exemples valides. Dans notre approche simple, deux précautions sont prises. D'une part, la longueur minimale de la sous-chaîne commune l est fixée à un nombre assez grand (dans nos expériences, $l = 7$ lettres), ce qui réduit le risque de réunir deux mots ne partageant aucun lien. D'autre part, comme cela a déjà observé [34], rechercher les variantes morphologiques au sein d'un même document maximise les chances que les deux mots soient issus d'une même thématique et donc d'un vocabulaire cohérent.

Une fois cette première phase accomplie, il nous est maintenant possible de vérifier si un couple de mots inconnus est en analogie avec une paire connue et de déduire ainsi si les deux mots inconnus sont en relation de dérivation ou de flexion. Dans le cadre de notre application, les mots dont on souhaite récupérer les variantes morphologiques sont ceux constituants les requêtes. Pour ce faire, chaque mot des requêtes est confronté à chaque mot de la collection ; si le couple ainsi formé est en analogie avec un des couples-exemples, il est alors utilisé pour l'extension de la requête. En pratique, pour des questions de rapidité, les règles d'analogies sont utilisées de manière

génératives : des mots sont produits à partir du terme de la requête en suivant les opérations de préfixation et suffixation indiquées dans les règles et ils sont conservés s'ils apparaissent dans l'index de la collection. L'apprentissage des règles se faisant hors-ligne, seule la recherche des variantes morphologiques des termes des requêtes dans l'index est faite en ligne. Cette recherche est d'une complexité en $\mathcal{O}(n)$ où n est le nombre de termes différents dans le corpus ; en pratique, dans les expériences reportées ci-après, cela prend quelques dixièmes de seconde sur un Pentium 1.5 GHz avec 512 Mo de RAM.

Ainsi, si l'on a la requête « *pollution des eaux souterraines* », la requête étendue finalement utilisée dans le SRI sera « *pollution des eaux souterraines polluants dépollution anti-pollution pollutions polluées polluent eau souterraine souterrains souterrain* ». Il est important de noter que, lors de l'extension, seuls les mots directement liés aux termes de la requêtes sont ajoutés ; les mots eux-mêmes liés aux extensions ne sont pas pris en compte. Cette absence volontaire de transitivité doit ainsi éviter de propager des erreurs (*vision* → *provision* → *provisions* → *provisionner* → *approvisionner* → *approvisionnement*...).

Dans les expériences présentées dans la section suivante, ce sont en moyenne 3 variantes morphologiques par mot plein de la requête qui sont ainsi ajoutées en extension. Aucun filtrage manuel des variantes n'est effectué et certaines extensions ne sont donc pas pertinentes. Il est cependant difficile d'effectuer une évaluation intrinsèque des extensions hors de leur cadre d'utilisation et cette évaluation ne préjugerait pas de l'influence de ces extensions sur le SRI (voir la discussion des résultats en section 4.5).

4 EXPÉRIENCES

Cette section présente les résultats obtenus par la méthode d'extension de requêtes décrite précédemment. Différentes expériences sont décrites : nous présentons tout d'abord les résultats de notre technique d'expansion sur deux collections françaises différentes (sous-section 4.1). Nous étudions ensuite l'intérêt de la prise en compte des préfixes dans notre technique de découverte de variantes (section 4.2). Nous nous intéressons en section 4.3 à l'influence de la longueur des requêtes sur les résultats et terminons en évaluant la portabilité de notre approche sur d'autres langues (sous-section 4.4).

Ces différentes expériences ont été menées sur la collection de données INIST composée de 30 requêtes et de 163 308 documents, résumés en français d'articles relevant de différentes disciplines scientifiques. Pour les expériences évaluant la portabilité de notre approche sur d'autres langues (section 4.4), nous utilisons la collection ELRA composée de 30 requêtes et 3511 documents issus de questions/réponses de la commission européenne et disponible en français, anglais, allemand, portugais, espagnol et italien.

Dans chacune de ces collections, les requêtes comportent plusieurs champs

(titre, question, informations complémentaires, concepts associés). Pour nous approcher d'un fonctionnement « grand-public » et donc utiliser des requêtes composées de peu de mots, les requêtes effectivement utilisées sont composées uniquement du contenu du champ titre, sauf en section 4.3 où nous étudions l'influence de la taille des requêtes sur notre technique d'extension. Le système de recherche d'information que nous utilisons pour ces expériences est LEMUR[1], paramétré de manière à adopter le fonctionnement du célèbre système OKAPI [29].

4.1 Résultats sur le français

Dans cette première expérience, nous utilisons la collection INIST et la collection ELRA en français, toutes les deux avec des requêtes courtes (champ sujet). L'apport de l'extension par variantes morphologiques est évalué en comparant les résultats obtenus avec et sans cette extension, classiquement mesurés en termes de précision et rappel à différents seuils (*i.e.* le rappel et précision calculés sur les n premiers documents trouvés, notés par la suite P(n) et R(n)), précision moyenne interpolée sur 11 points (IAP), R-précision et précision moyenne non-interpolée (MAP).

À titre de comparaison, nous présentons également les résultats obtenus par deux systèmes standard manipulant des informations morphologiques usuellement utilisés en RI pour leur robustesse et leur disponiblité : un *stemmer* du français (développé par J. Savoy [31] et s'appuyant sur un ensemble de règles fixées de désuffixation), et un lemmatiseur du français (l'étiqueteur TREETAGGER[2]). Ces deux techniques fonctionnent différemment de notre approche puisque la variation morphologique est prise en compte dès la phase d'indexation : les mots sont ramenés à une forme canonique (*stem* ou lemme) avant d'être inclus dans l'index. Des techniques d'expansion automatique de requêtes existent (par exemple celles basées sur le *pseudo-relevance feedbak* [30]) mais ne portent pas spécialement sur la variation morphologique ; nous ne les présentons pas ici.

Les tableaux 1 et 2 récapitulent les résultats respectivement obtenus pour la collection INIST et ELRA ; les chiffres jugés statistiquement non significatifs par un *paired t-test* [14] (avec une valeur $p < 0.05$) apparaissent en petites italiques. La taille moyenne des requêtes (|Q|) est également indiquée, calculée en nombre de mots (y compris les mots outils).

Il ressort de ces deux tableaux que notre approche obtient de très bons résultats pour chacune des mesures adoptées, tous statistiquement significatifs à deux exceptions près dans la collection ELRA. Pour la plupart des mesures, l'extension de requêtes est notamment plus performante que le *stemming* ou la lemmatisation, mais aussi plus stable comme l'atteste certains résultats jugés non statistiquement significatifs de ces deux techniques. Les

[1] LEMUR est disponible à l'URL http://www.lemurproject.org

[2] TREETAGGER est disponible à l'URL http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

	Sans extension	Avec extension (amélioration %)	*Stemming* (amélioration %)	Lemmatisation (amélioration %)
|Q|	5.3	16.03	5.2	5.17
MAP	14.85	18.45 (+24.29%)	*17.31 (+16.63%)*	17.82 (+20.07%)
IAP	16.89	19.93 (+17.97%)	*18.85 (+11.57%)*	19.72 (+16.73%)
R-Prec	17.99	21.63 (+20.24%)	*19.88 (+10.53%)*	*19.71 (+9.56%)*
P(10)	34.33	38.67 (+12.62%)	*36.67 (+6.80%)*	39.67 (+15.53%)
P(20)	27.83	31.83 (+14.37%)	*29.00 (+4.19%)*	31.6 (+13.77%)
P(50)	18.33	21.27 (+16.00%)	*20.13 (+9.82%)*	*20.87 (+13.82)*
P(100)	12.23	14.80 (+20.98%)	15.23 (+24.52%)	14.97 (+22.34%)
P(200)	8.02	9.73 (+21.41%)	9.77 (+21.82%)	9.5 (+18.50%)
P(500)	3.88	4.80 (+23.71%)	4.55 (+17.18%)	4.47 (+15.29%)
P(1 000)	2.21	2.68 (+21.30%)	2.53 (+14.80%)	2.48 (+12.39%)
P(2 000)	1.29	1.50 (+16.45%)	1.40 (+8.68%)	1.39 (+8.29%)
P(5 000)	0.56	0.67 (+20.38%)	0.63 (+13.47%)	0.64 (+15.14%)
R(10)	8.00	8.99 (+12.36%)	*8.45 (+5.64%)*	*9.04 (+13.02%)*
R(20)	12.33	14.50 (+17.59%)	*12.81 (+3.90%)*	*13.62 (+10.48%)*
R(50)	19.65	24.07 (+22.47%)	*20.78 (+5.74%)*	*21.56 (+9.71%)*
R(100)	26.85	32.87 (+22.41%)	31.32 (+16.64%)	31.58 (+17.59%)
R(200)	34.38	42.70 (+24.21%)	41.05 (+19.43%)	40.92 (+19.04%)
R(500)	43.09	53.83 (+24.92%)	49.31 (+14.43%)	49.35 (+14.54%)
R(1 000)	48.43	59.45 (+22.74%)	55.27 (+14.12%)	55.03 (+13.62%)
R(2 000)	55.35	65.85 (+18.99%)	61.08 (+10.36%)	61.20 (+10.57%)
R(5 000)	59.32	72.20 (+21.71%)	67.22 (+13.31%)	68.20 (+14.96%)

TAB. 1 – Performances de l'extension de requête sur la collection INIST

résultats semblent largement dépendants de la collection utilisée. À ce titre, notre approche semble néanmoins assez robuste, contrairement à la lemmatisation qui n'apporte aucune amélioration pour la collection ELRA, ou au *stemming* dont l'effet n'est sensible (et restreint) qu'après les 500 premiers documents. Il est également intéressant de noter que l'amélioration des résultats est également distribuée sur tous les seuils de mesures (de 10 à 5 000 documents), quelle que soit la collection utilisée. Cela signifie que l'amélioration n'est pas due à une simple réorganisation des documents en tête de liste mais plutôt à la découverte de documents pertinents qui n'auraient pas été ramenés par le SRI sans les extensions de requêtes.

4.2 Influence de la prise en compte des préfixes

À l'inverse de beaucoup de techniques prennant en compte la variation morphologique en RI, notre approche prend en compte l'opération de préfixation. Si cette opération semble utile dans certains cas, on peut craindre

	Sans extension	Avec extension (amélioration %)	*Stemming* (amélioration %)	Lemmatisation (amélioration %)
|Q|	2.83	9.07	2.8	2.83
MAP	42.27	47.47 (+12.30%)	*42.63 (+0.87%)*	*41.32 (-2.24%)*
IAP	43.15	48.00 (+11.22%)	*44.05 (+2.11%)*	*42.58 (-1.3%)*
R-Prec	44.22	48.54 (+9.76%)	*45.1 (+2.1%)*	*42.89 (-3.01%)*
P(10)	56.67	*58.33 (+2.94%)*	*54.67 (-3.53%)*	*56.33 (-0.59%)*
P(20)	48.67	51.67 (+6.16%)	*48.17 (-1.03%)*	*49 (+0.68%)*
P(50)	30.33	34.6 (+14.07%)	*30.4 (+0.22%)*	*29.13 (-3.96%)*
P(100)	19.7	23 (+16.75%)	*20.1 (+2.38%)*	*18.1 (-8.12%)*
P(200)	11.42	12.87 (+12.7%)	*11.58 (+2.81%)*	*10.65 (-6.72%)*
P(500)	4.84	5.68 (+17.19%)	4.95 (+5.99%)	*4.64 (-4.26%)*
P(1 000)	2.45	2.89 (+18.12%)	2.52 (+6.64%)	*2.24 (-4.22%)*
P(2 000)	1.22	1.45 (+18.53%)	1.27 (+7.49%)	*1.17 (-4.22%)*
R(10)	22.36	*23.49 (+5.08%)*	*22.27 (-0.37%)*	*22.44 (+0.37%)*
R(20)	35.55	37.93 (+6.69%)	*37.73 (+6.14%)*	*37.39 (+5.19%)*
R(50)	49.30	54.84 (+11.24%)	*51.41 (+4.28%)*	*49.98 (+1.39%)*
R(100)	60.04	66.39 (+10.59%)	*62.12 (+3.64%)*	*58.27 (-2.95%)*
R(200)	68.52	73.39 (+7.12%)	*70.84 (+4.08%)*	*66.96 (-2.27%)*
R(500)	72.50	81.48 (+12.40%)	76.37 (+7.29%)	*72.53 (+0.05%)*
R(1 000)	73.42	82.87 (+12.87%)	77.56 (+7.57%)	*73.83 (+0.55%)*
R(2 000)	73.42	83.24 (+13.38%)	78.19 (+8.43%)	*73.83 (+0.55%)*

TAB. 2 – Performances de l'extension de requête sur la collection ELRA

qu'elle apporte plus de bruit (couples de mots non liés sémantiquement ou faiblement liés) que de couples utiles pour l'expansion. En effet, l'ajout d'un préfixe en français, signe d'une dérivation, induit souvent une modification de sens plus importante que pour la suffixation (le plus souvent le signe d'une flexion). Pour vérifier l'intérêt de la préfixation dans notre cadre de RI, nous répétons donc les deux expériences précédentes en n'autorisant que les extensions trouvées par changement, ajout ou suppression de suffixe. Le tableau 3 présente les résultats obtenus sur les collections INIST et ELRA français ; comme précédemment, l'amélioration relative est calculée à partir des résultats sans aucune extension. À titre de comparaison, on répète les résultats obtenus en prenant en compte les préfixes et les suffixes.

Les résultats de ces expériences montrent tout d'abord que la prise en compte de la préfixation n'apporte que quelques extensions (un peu plus d'un terme par requête). Cependant, ces extensions apportent un gain léger mais constant de performances à tous les seuils de mesures pour les deux collections. L'opération de préfixation semble donc être bénéfique à notre système d'extension même si la plus grande partie des gains de performances est due à des variantes issues de changement de suffixes.

	INIST		ELRA	
	Préfixes et suffixes	Suffixes seulement	Préfixes et suffixes	Suffixes seulement
|Q|	16.03	14.77	9.07	7.93
MAP	18.45 (+24.29%)	17.87 (+20.34%)	47.47 (+12.30%)	47.4 (+12.14%)
IAP	19.93 (+17.97%)	19.59 (+16.06%)	48.00 (+11.22%)	47.99 (+11.21%)
R-Prec	21.63 (+20.24%)	20.79 (+15.58%)	48.54 (+9.76%)	48.74 (+10.21%)
P(10)	38.67 (+12.62%)	38.67 (+12.62%)	*58.33 (+2.94%)*	*58.33 (+2.94%)*
P(20)	31.83 (+14.37%)	16.67 (+13.77%)	51.67 (+6.16%)	51.67 (+6.16%)
P(50)	21.27 (+16.00%)	20.93 (+14.23%)	34.6 (+14.07%)	34.67 (+14.29%)
P(100)	14.80 (+20.98%)	14.47 (+18.26%)	23 (+16.75%)	23 (+16.75%)
P(200)	9.73 (+21.41%)	9.57 (+19.33%)	12.87 (+12.7%)	12.88 (+12.85%)
P(500)	4.80 (+23.71%)	4.69 (+20.58%)	5.68 (+17.19%)	5.66 (+16.78%)
P(1 000)	2.68 (+21.30%)	2.6 (+17.83%)	2.89 (+18.12%)	2.88 (+17.71%)
P(2 000)	1.50 (+16.45%)	1.46 (+13.73%)	1.45 (+18.53%)	1.44 (+17.98%)
P(5 000)	0.67 (+20.38%)	0.66 (+17.16%)	0.58 (+18.53%)	0.58 (+17.98%)
R(10)	8.99 (+12.36%)	8.76 (+9.45%)	*23.49 (+5.08%)*	*23.43 (+4.81%)*
R(20)	14.50 (+17.59%)	14.24 (+15.54%)	37.93 (+6.69%)	38.02 (+6.97%)
R(50)	24.07 (+22.47%)	22.88 (+16.46%)	54.84 (+11.24%)	54.92 (+11.4%)
R(100)	32.87 (+22.41%)	31.99 (+19.12%)	66.39 (+10.59%)	66.35 (+10.52%)
R(200)	42.70 (+24.21%)	41.95 (+22.02%)	73.39 (+7.12%)	73.42 (+7.15%)
R(500)	53.83 (+24.92%)	52.28 (21.33%)	81.48 (+12.40%)	81.24 (+12.06%)
R(1 000)	59.45 (+22.74%)	57.84 (+19.43%)	82.87 (+12.87%)	82.63 (+12.55%)
R(2 000)	65.85 (+18.99%)	64.10 (+15.81%)	83.24 (+13.38%)	82.91 (+12.93%)
R(5 000)	72.20 (+21.71%)	70.17 (+18.29%)	83.24 (+13.38%)	82.91 (+12.93%)

TAB. 3 – Performances de l'extension de requête avec et sans préfixation

4.3 Influence de la taille des requêtes

Pour mesurer l'influence de la taille de la requête sur notre approche, nous répétons l'expérience précédente en utilisant cette fois-ci les autres champs des requêtes INIST pour composer des requêtes de plus en plus longues. Plus précisément, ce sont les champs *concept* qui sont ajoutés un par un à la requête (initialement composée du seul champ sujet). La figure 1 présente les résultats obtenus selon la taille, mesurée en nombre de mots avant toute extension, des requêtes ainsi constituées. La performance du SRI est mesurée par la précision moyenne non-interpolée.

Quelle que soit la taille de la requête et la technique adoptée, il ressort de ces résultats l'intérêt de gérer les variantes morphologiques. Parmi les trois techniques à l'épreuve, notre approche d'extensions morphologiques reste la plus performante, devant le *stemming* et la lemmatisation. Il est également intéressant de constater que contrairement à ce qui est parfois avancé, la

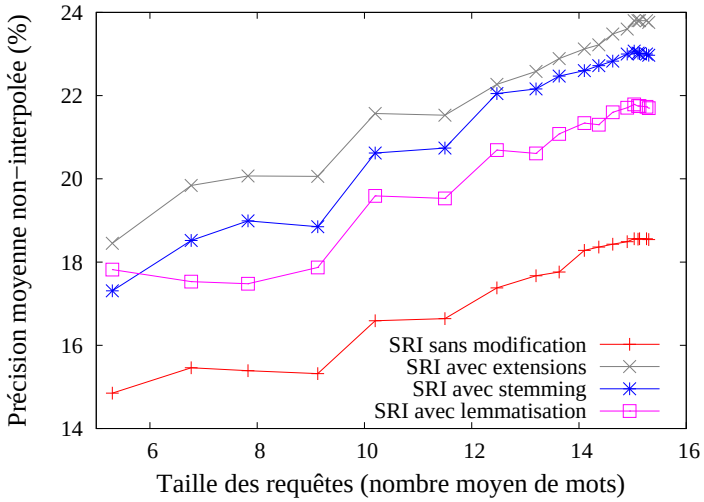

FIG. 1 – Évolution de la précision selon la taille de la requête

prise en compte de la morphologie apporte un gain de performance constant, même pour des requêtes longues.

4.4 Application à d'autres langues

Notre approche, n'utilisant aucune connaissance externe, se veut portable ; elle doit donc être directement utilisable pour n'importe quelle langue dont la morphologie se fait principalement par préfixation et suffixation. Pour vérifier la portée de cette assertion, nous présentons ci-dessous dans le tableau 4 les résultats obtenus sur la collection ELRA pour l'allemand, l'anglais, l'espagnol, le français, l'italien et le portugais. Pour chacune de ces langues, nous indiquons la variation (en pourcentage) par rapport à la même recherche utilisant les requêtes sans extension.

Les résultats sont tous très positifs puisque l'extension de requêtes apporte un gain de performance de 10 à 20% selon les langues et les mesures. Comme pour le français pour les expériences précédentes, l'amélioration se fait sentir à tous les seuils de mesures de précision et de rappel ; cependant, pour les seuils bas (10 à 50 documents), les résultats sont très variables d'un requête à une autre, ce que traduit le fait qu'ils ne soient pas jugés statistiquement significatifs.

Là encore, quelques résultats allant à l'encontre de ce qui est parfois avancé par certains auteurs se font jour. Tout d'abord, l'anglais, réputée de morphologie pauvre, bénéficie plus de l'extension de requêtes par variantes morphologiques que des langues à la morphologie plus riche (espagnol, italien...).

	Langues					
	Allemand	Anglais	Espagnol	Français	Italien	Portugais
MAP	+16.25%	+17.52%	+10.03%	+11.89%	+10.45%	+9.69%
IAP	+15.93%	+16.66%	+8.70%	+10.99%	+9.79%	+9.25%
R-Prec	*+3.03%*	+10.23%	+7.97%	+9.43%	+10.23%	*+6.20%*
P(10)	*+10.68%*	+7.03%	*0%*	*+3.53%*	*+2.54%*	*0%*
P(20)	*+8.33%*	*+3.62%*	+7.41%	+6.85%	+11.15%	*+4.38%*
P(50)	*+6.69%*	+8.23%	+13.40%	+13.85%	+13.48%	+8.31%
P(100)	+9.54%	+14.31%	+16.76%	+16.24%	+18.98%	+14.24%
P(200)	+12.61%	+19.63%	+14.15%	+12.70%	+18.70%	+19.27%
P(500)	+13.18%	+20.49%	+18.13%	+17.19%	+18.94%	+23.35%
P(1 000)	+12.97%	+21.60%	+20.32%	+18.26%	+22.13%	+24.64%
P(2 000)	+12.33%	+19.94%	+20.70%	+18.66%	+22.71%	+24.85%
R(10)	*+6.82%*	+2.90%	*+1.88%*	*+5.43%*	*-0.67%*	*-0.47%*
R(20)	*+5.95%*	*+3.27%*	+7.40%	+7.36%	+7.82%	*+7.55%*
R(50)	*+11.12%*	+8.48%	+7.72%	+10.82%	+7.37%	*+6.21%*
R(100)	+11.87%	+13.23%	+10.14%	+10.11%	+8.93%	+9.39%
R(200)	+18.23%	+19.43%	+7.40%	+6.92%	+9.77%	+14.18%
R(500)	+16.45%	+21.68%	+14.49%	+12.69%	+14.31%	+17.71%
R(1 000)	+18.15%	+20.93%	+17.38%	+13.20%	+18.35%	+19.23%
R(2 000)	+16.34%	+17.77%	+17.54%	+13.70%	+18.97	+19.26%

TAB. 4 – Performances de l'extension de requête sur différentes langues

Autre résultat intéressant, l'allemand est la langue bénéficiant le plus de notre technique d'extensions, sans doute par le fait que notre approche capture des cas d'agglutination fréquente dans cette langue (par exemple, le couple *Menschenrechte* ↔ *Menschenrechtsorganisation*).

4.5 Discussion des résultats

De ces expériences, il est difficile de tirer des conclusions allant dans le sens de celles parfois avancées dans ce domaine, si ce n'est que la prise en compte de la variation morphologique par notre approche améliore bien les performances des SRI. En revanche, les différences de richesse morphologique des langues tendant à opposer l'anglais à l'espagnol ou au portugais ne semblent pas avoir d'influence directe dans nos expériences, contrairement à d'autres expériences [1, par exemple]. D'autre part, la taille des requêtes ne semblent pas non plus influencer dans un sens ou l'autre les résultats ; les améliorations sont constantes et de même ordre de grandeur pour des requêtes allant de 5 à 15 mots.

Si les résultats sont tous positifs et constants au sein d'une collection, on constate néanmoins une variation importante de performances entre les col-

lections pour une même langue. Il semble donc que ce type d'approche prenant en compte la variation morphologique soit en fait sensible à la collection plus qu'à la langue. Il faut à ce titre souligner quelques particularités de la collection multilingue ELRA qui peuvent éclairer les résultats obtenus pour le français, mais aussi pour les autres langues, lors de nos expérimentations. Tout d'abord, la collection est très petite (environ 3 500 documents contre 163 308 pour la collection INIST), les requêtes sont beaucoup plus courtes (il y a donc moins de mots pleins auxquels il est possible d'ajouter des variantes), et enfin les types de documents (débats parlementaires contre résumés articles scientifiques) doivent certainement avoir un impact sur la productivité morphologique.

Notre méthode pour enrichir les termes de la requête à l'aide de leurs variantes morphologiques n'est pas exempte d'erreurs. Certains termes morphologiquement liés ne sont pas trouvés, et, ce qui est plus préjudiciable pour l'extension de requêtes, des termes non pertinents sont parfois trouvés. Dans ce dernier type d'erreur, il faut distinguer plusieurs cas. Tout d'abord, certains mots trouvés n'entretiennent pas de liens sémantiques avec le terme original, le lien morphologique étant fortuit (*pondre* ↔ *répondre*) ou pertinent d'un point de vue diachronique mais plus en œuvre dans la langue actuelle, comme par exemple *composition* ↔ *exposition*, ou *ordinateur* ↔ *ordination*.

Ensuite, certains termes polysémiques provoquent des erreurs difficiles à éviter, mettant en valeur l'intérêt qu'il y aurait à utiliser des outils de désambiguïsation de sens. Ainsi, les deux mots *production* et *reproduction* trouvés comme étant liés, le sont dans les phrases *production des résultats* et *reproduction des résultats*, mais pas avec *la reproduction chez les poissons*. Rappelons à ce titre que l'absence volontaire de transitivité dans nos extensions, c'est-à-dire le fait que l'on n'ajoute pas de mots liés aux extensions permet de limiter la portée de ces erreurs. À ce titre, l'extension de requête apparaît comme un cadre plus souple que les techniques fonctionnant par conflation dans lesquelles *production* et *reproduction* ou *départ* et *département* et tous leurs dérivés seraient indissociablement réunis sous un seul identifiant.

Enfin, dans certains cas, les deux mots sont bien reliés sémantiquement et semblent bien correspondre au sens utilisé dans la requête, et pourtant l'ajout du terme lié dégrade les performances du SRI ; par exemple, dans la collection INIST l'extension de la requête *impact sur l'environnement des moteurs diesel* avec le terme *moteur* fait chuter la qualité des résultats proposés.

5 CONCLUSION ET PERSPECTIVES

Nous avons proposé dans cet article une technique simple permettant de détecter automatiquement des variantes morphologiques au sein de textes en s'appuyant sur la construction d'analogies. Cette technique, qui repose sur l'aprentissage de schémas de préfixation et de suffixation, a été utilisée pour étendre des requêtes avec les variantes morphologiques des termes

initialement présents dans ces requêtes. L'objectif de ce travail était donc d'améliorer les performances des SRI en utilisant ces connaissances morphologiques. À ce titre, les résultats obtenus par notre approche non-supervisée sont très satisfaisants et se comparent avantageusement aux outils supervisés testés (*stemmer* à base de règles et lemmatiseur). Nos expériences avec ces autres outils morphologiques confirment d'ailleurs les résultats mitigés qui ressortent de l'état de l'art ; en effet, si en moyenne le *stemming* ou la lemmatisation apportent un gain réel, leurs résultats, contrairement à notre approche, ne sont pas toujours statistiquement significatifs, ce qui indique un important manque de stabilité des résultats requête par requête. L'amélioration des performances amenée par notre méthode et constatée sur les deux collections de textes en français est assez importante (entre 12 et 24%) et constante. Sur les collections en français, l'opération de préfixation n'apporte que peu de variantes, mais elles sont cependant pertinentes et apportent un gain léger de performances. De plus, on a montré que l'aspect non-supervisé de notre technique d'extension, point essentiel pour avoir sa portabilité, permettait de l'appliquer à diverses langues. Là encore, les expérimentations menées montrent le bien-fondé de notre approche aussi bien pour l'allemand, l'anglais, l'espagnol, le français, l'italien et le portugais, avec des gains de performances importants et assez homogènes d'une langue à l'autre.

Beaucoup de perspectives sont envisagées à la suite de ce travail. D'un point de vue technique tout d'abord, la mesure Sim à la base de notre technique d'acquisition peut être revue pour, d'une part, prendre en compte les infixes, et d'autre part, gérer plus naturellement les modifications légères de racines et ainsi permettre d'accepter des analogies comme *majeur* : *majorité* $\doteq$ *stupide* : *stupidité*. Concernant l'utilisation des variantes en RI, la contrainte forte interdisant d'ajouter les mots morphologiquement liés aux termes déjà ajoutés à la requête pourrait être assouplie. À l'inverse, plutôt que d'ajouter tous les mots supposés liés à un terme de la requête comme c'est actuellement le cas, seuls certains jugés plus pertinents pourraient être conservés. Pour ce faire, la décision d'ajout à la requête pourrait être basée sur la confiance que l'on a dans l'analogie (évaluée en fonction de sa productivité par exemple) et sur l'importance (le poids) du terme de la requête auquel il est directement ou indirectement lié. D'un point de vue applicatif, les résultats sur les langues étudiées demandent à être vérifiés et consolidés sur d'autres collections, et étendus à d'autres langues.

Enfin, dans un cadre de RI translingue, l'utilisation d'une approche similaire basée sur l'analogie pour la traduction de termes spécialisés est à l'étude. Le principe serait de chercher des règles analogiques pour capturer les régularités de traduction existant dans certains domaines entre certaines langues (voir par exemple [22] dans le domaine biomédical)

RÉFÉRENCES

[1] Avi Arampatzis, Theo P. Van Der Weide, Cornelis H. A. Koster et Patrick Van Bommel. *Linguistically Motivated Information Retrieval*, volume 69, pages 201–222. M. Dekker, New York, États-Unis, 2000.

[2] Matthew W. Bilotti, Boris Katz et Jimmy Lin. What Works Better for Question Answering : Stemming or Morphological Query Expansion ? In *Proceedings of the Information Retrieval for Question Answering (IR4QA) Workshop at SIGIR 2004*, 2004.

[3] Vincent Claveau et Marie-Claude L'Homme. Apprentissage par analogie de liens sémantiques entre dérivés morphologiques. In *Actes de la conférence de Terminologie et Intelligence Artificielle, TIA'05*, Rouen, France, 2005.

[4] Vincent Claveau et Marie-Claude L'Homme. Structuring terminology by analogy machine learning. In *Proceedings of the International conference on Terminology and Knowledge Engineering, TKE'05*, Copenhague, Danemark, 2005.

[5] Claude de Loupy. *Évaluation de l'apport de connaissances linguistiques en desambigüisation semantique et recherche documentaire*. Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse, 2000.

[6] George E. Freund et Peter Willett. Online Identification of Word Variants and Arbitrary Truncation Searching Using a String Similarity Measure. *Information Technology : Research and Development*, 1 :177–187, 1982.

[7] Michael Fuller et Justin Zobel. Conflation-Based Comparison of Stemming Algorithms. In *Proceedings of the 3rd Australian Document Computing Symposium*, Sydney, Australie, 1998.

[8] Eric Gaussier. Unsupervised Learning of Derivational Morphology from Inflectional Corpora. In *Proceedings of Workshop on Unsupervised Methods in Natural Language Learning, 37th Annual Meeting of the Association for Computational Linguistics*, Maryland, États-Unis, 1999.

[9] Eric Gaussier, Grégory Grefenstette, David Hull et Claude Roux. Recherche d'information en Français et traitement automatique des langues. *Traitement automatique des langues*, 41(2) :473–493, 2000.

[10] John A. Goldsmith. Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*, 27(2) :153–198, 2001.

[11] John A. Goldsmith, Derrick Higgins et Svetlana Soglasnova. Automatic Language-Specific Stemming in Information Retrieval. In *Cross-Language Information Retrieval and Evaluation : Proceedings of the CLEF 2000 workshop*, pages 273–284. Springer Verlag, 2001.

[12] Donna Harman. How Effective is Suffixing ? *Journal of the American Society for Information Science*, 42(1) :7–15, 1991.

[13] Nabil Hathout. Analogies morpho-synonimiques. une méthode d'acquisition automatique de liens morphologiques à partir d'un dictionnaire de synonymes. In *Actes de la 8^{e} conférence Traitement Automatique du Langage Naturel, TALN'01*, Tours, France, 2001.

[14] David Hull. Using Statistical Testing in the Evaluation of Retrieval Experiments. In *Proceedings of the 16th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'93*, Pittsburgh, États-Unis, 1993.

[15] David Hull. Stemming Algorithms - A Case Study for Detailed Evaluation. *Journal of the American Society of Information Science*, 47(1) :70–84, 1996.

[16] David Hull et Grégory Grefenstette. A Detailed Analysis of English Stemming Algorithms. Technical Report, Xerox Research Centre Europe, Meylan, France, 1996.

[17] Robert Krovetz. Viewing Morphology as an Inference Process. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, États-Unis, 1993.

[18] Mikko Kurimo, Mathias Creutz et Krista Lagus, éditeurs. *Unsupervised Segmentation of Words into Morphemes Challenge*, Venise, Italie, Avril 2005. Associé au second Workshop Challenge Pascal.

[19] Martin Lennon, David S. Pierce, Brian D. Tarry et Peter Willett. An Evaluation of some Conflation Algorithms for Information Retrieval. *Journal of Information Science*, 3(1) :177–183, 1981.

[20] Yves Lepage. *De l'analogie ; rendant compte de la communication en linguistique*. Thèse d'habilitation (HDR), Université de Grenoble 1, Grenoble, France, 2003.

[21] Julie Beth Lovins. Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics*, 1 :22–31, 1968.

[22] Kornél Markó, Stefan Schulz, Olena Medelyan et Udo Hahn. Bootstrapping dictionaries for cross-language information retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*, Salvador, Brésil, 2005.

[23] Igor Mel'čuk. *Cours de morphologie générale*, volume 1-5. Presses de l'Université de Montréal/CNRS Éditions, Montréal/Paris, 1993-2000.

[24] Fabienne Moreau et Pascale Sébillot. Contributions des techniques du traitement automatique des langues à la recherche d'information. Rapport de recherche 1690, IRISA, 2005.

[25] Isabelle Moulinier, J. Andrew McCulloh et Elizabeth Lund. West Group at CLEF 2000 : Non-English Monolingual Retrieval. In *Proceedings of the Workshop of Cross-Language Evaluation Forum, CLEF 2000*, Lisbonne, Portugal, 2000.

[26] Douglas W. Oard, Gina-Anne Levow et Clara I. Cabezas. Clef Experiments at Maryland : Statistical Stemming and Backoff Translation. In *Cross-Language Information Retrieval and Evaluation : Proceedings of the CLEF 2000 workshop*, pages 176–187. C. Peters, Ed. : Springer Verlag, 2001.

[27] Mirko Popovic et Peter Willett. The Effectiveness of Stemming for Natural-Language Access to Slovene Textual Data. *Journal of the American Society for Information Science*, 43(5) :384–390, 1992.

[28] Martin Porter. An Algorithm for Suffix Stripping. *Program*, 14(3) :130–137, 1980.

[29] Stephen E. Robertson, Steve Walker et Micheline Hancock-Beaulieu. Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive. In *Proceedings of the 7th Text Retrieval Conference, TREC-7*, pages 199–210, 1998.

[30] Joseph J. Rocchio. Relevance Feedback in Information Retrieval. In Gerard Salton, éditeur, *The SMART Retrieval System : Experiments in Automatic Document Processing*, pages 313–323. Prentice-Hall, Englewood Cliffs, 1971.

[31] Jacques Savoy. A stemming procedure and stopword list for general french corpora. *Journal of the American Society for Information Science*, 50(10), 1999.

[32] Jacques Savoy. Morphologie et Recherche d'Information. Rapport technique, Institut interfacultaire d'informatique, Université de Neuchâtel, 2002.

[33] Jesus Vilares Ferro, Mario Barcala et Miguel A. Alonso. Using Syntactic Dependency-Pairs Conflation to Improve Retrieval Performance in Spanish. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics, CICLING*, Mexico, Mexique, 2002.

[34] Jinxi Xu et Bruce W. Croft. Corpus-based stemming using cooccurrences of words variants. *ACM Transactions on Information Systems*, 16(1) :61–81, 1998.

[35] Jinxi Xu et Bruce W. Croft. Improving the Effectiveness of Information Retrieval with Local Context Analysis. *ACM Transactions on Information Systems*, 18(1) :79–112, 2000.