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}

Catégorie : chercheur.

RÉSUMÉ. *Cet article s'intéresse au problème de la formulation différente d'une même idée, d'un même concept, en recherche d'information à travers la prise en compte du phénomène de la variation morphologique. 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 déjà réalisés dans ce domaine, la technique proposée présente la particularité de ne nécessiter aucunes ressources ni connaissances externes, et d'être applicable par conséquent à une grande variété de langues. Les évaluations de cette approche réalisées sur plusieurs collections de documents, sur 6 différentes 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.*

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.*

MOTS-CLÉS : *Variation morphologique, extension de requêtes, apprentissage automatique par analogie.*

KEYWORDS: *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) 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 de mots, les SRI se retrouvent rapidement confrontés à deux problèmes duaux liés à la complexité du langage naturel. Le premier est lié à la polysémie : un même terme peut référer à des concepts différents, ce qui entraîne potentiellement la récupération par le SRI de documents non pertinents. La seconde difficulté est due à la possibilité de formuler différemment un même concept. Un document pertinent peut ainsi contenir des termes différents de ceux de la requête bien que « sémantiquement proches ».

Cet article présente une approche pour contourner en partie ce second problème par la prise en compte de la variation morphologique. C'est en effet ce type de variation qui empêche par exemple une requête contenant le terme *déménager* d'être apparié avec un document contenant *déménagement*. La connaissance des relations morphologiques permettrait ainsi aux SRI de retrouver davantage de documents pertinents. Ce 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 à ce problème de portabilité, nous présentons une approche simple et efficace avec les hypothèses suivantes :

– le système ne doit nécessiter aucune connaissance ou donnée externes ;
– le système doit être complètement automatique ;
– le système doit pouvoir s'appliquer directement à diverses langues.

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 utilisation pour l'extension de requêtes. 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

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 dis-

tingue usuellement plusieurs types de relations morphologiques [MEL 00] ; 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. 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* ↔ *défaire*) et/ou de catégories grammaticales (*décider* ↔ *décision*). Ainsi, des termes qui dérivent du même lemme ou de la même racine présupposent donc généralement un sens proche. La variation morphologique constitue un type particulier de variation sémantique qu'il est intéressant de capturer en RI.

2.2. *Travaux connexes*

Beaucoup de travaux ont été menés pour prendre en compte la variation morphologiques au sein des SRI. Un état de l'art très complet peut être trouvé dans [MOR 05], nous n'en donnons ici que les grandes lignes et nous concentrons sur les travaux les plus proches des nôtres.

Deux approches différentes permettent la prise en compte de la variation morphologique au sein d'un SRI. Dans la première, la conflation, les différentes variantes morphologiques possibles d'un mot sont ramenées à une seule et même forme, racine ou lemme, et l'appariement des documents et de la requête se fait sur la base de cette forme canonique. La seconde approche est l'expansion : les termes des requêtes 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 flexionnelle et dérivationnelle d'un mot autour d'un *stem* (*i.e.* une pseudo-racine). Les techniques utilisées pour ce faire reposent généralement sur une liste d'affixes de la langue considérée et sur un ensemble de règles de désuffixation construites *a priori* [POR 80, LOV 68] qui permettent, étant donné un mot de trouver son *stem*. L'impact du *stemming* sur les performances des SRI est mitigé : les expériences montrent des améliorations faibles des résultats, aussi bien en indexation qu'en expansion [FUL 98, LOU 00], et variables [KRO 93, HUL 96]. D'autres expériences en RI ont tenté de prendre en compte la variation morphologique à l'aide d'outils plus « sophistiqués » que les *stemmers*, comme les lemmatiseurs et des outils d'analyses dérivationnelles. Là encore un gain de performances variable selon les langues et les expériences mais globalement positif est constaté [VIL 02, SAV 02, *inter alia*].

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. Citons tout d'abord l'approche adoptée par Xu et Croft [XU 00] qui s'appuie 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* [FRE 82] peuvent également être vus comme des techniques permettant de contourner le problème de la variation morphologique. Malheureusement, cette approche un peu brutale n'apporte en pratique pas ou peu d'amélioration des résultats [SAV 02].

Un travail très proche du nôtre est celui de Gaussier *et al.* [GAU 00] dans lequel est proposée une méthode d'apprentissage non supervisée des suffixes et des opérations de suffixation à partir de lexiques flexionnels de la langue 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 relations morphologiques intéressantes (*e.g. foudre* $\leftrightarrow$ *parafoudre*, *cancéreux* $\leftrightarrow$ *anticancéreux*). D'autre part, les travaux de Gaussier *et al.* tirent parti d'informations catégorielles (nom, verbe, adjectif...) sur les mots manipulés, ce qui implique soit l'emploi de données externes, soit d'outils dédiés. Cela 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 al* fonctionne par conflation alors que nous utilisons les variantes morphologiques uniquement en expansion. Cette dernière méthode nous paraît en effet plus flexible et moins brutale pour prendre en compte les variations morphologiques (voir notamment [BIL 04, XU 00] sur ce sujet).

3. Acquisition des variantes morphologiques en RI

Dans cette section, nous présentons tout d'abord la technique d'acquisition de variantes morphologiques utilisée. Nous détaillons ensuite 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 [CLA 05]. 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 [HAT 01, LEP 03] : 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 :

$$\mathrm{Sim}(m_1\text{-}m_2, m_3\text{-}m_4) = 1 \quad \text{si} \quad \begin{cases} m_1 = \mathrm{lcss}(m_1, m_2) +_{pre} \mathrm{Pre}_1 +_{suf} \mathrm{Suf}_1, \text{ et} \\ m_2 = \mathrm{lcss}(m_1, m_2) +_{pre} \mathrm{Pre}_2 +_{suf} \mathrm{Suf}_2, \text{ et} \\ m_3 = \mathrm{lcss}(m_3, m_4) +_{pre} \mathrm{Pre}_1 +_{suf} \mathrm{Suf}_1, \text{ et} \\ m_4 = \mathrm{lcss}(m_3, m_4) +_{pre} \mathrm{Pre}_2 +_{suf} \mathrm{Suf}_2 \end{cases}$$

$\mathrm{Sim}(m_1\text{-}m_2, m_3\text{-}m_4) = 0$ sinon

où $\mathrm{Pre_i}$ et $\mathrm{Suf_i}$ sont des chaînes de caractères quelconques. Si $\mathrm{Sim}(m_1\text{-}m_2, m_3\text{-}m_4) = 1$, 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. 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* $\leftrightarrow$ *réinstallation* n'est pas stocké en tant que tel, mais on conserve la règle : $m_2 = m_1 -_{pre}$ *dés* $+_{pre}$ *ré* $-_{suf}$ *er* $+_{suf}$ *ation*
Montrer l'analogie *déshydrater* : *réhydratation* $\doteq$ *désinstaller* : *réinstallation* revient alors simplement à tester que *déshydrater* $\leftrightarrow$ *réhydratation* vérifie la règle précédente.

Cette approche simple par analogie permet 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 (voir [CLA 05] 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 partageant ; 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 chiffre assez grand (dans nos expériences, $l = 7$ lettres), ce qui réduit le risque de réunir deux mots ne partageant aucun lien. D'autre part, 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 ; en pratique, dans les expériences reportées ci-après, cela prend quelques dixièmes de seconde.

Ainsi, pour une 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*...).

4. Expériences

Cette section présente les résultats obtenus par la méthode d'extension de requêtes décrite précédemment. Ces expériences ont été menées principalement sur la collection de données INIST composées de 30 requêtes et de 163 308 documents, résumés français d'articles de différentes disciplines scientifiques. Pour les expériences évaluant la portabilité de notre approche sur d'autres langues (section 4.3), nous utilisons la collection ELRA composée de 30 requêtes et 3511 documents issus de questions/réponses de la commission européenne. Pour 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 4.2 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 [ROB 98].

4.1. *Premiers résultats*

Dans cette première expérience, nous utilisons la collection INIST 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 sur les n premiers documents trouvés (P(n) et R(n) par la suite), 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 utilisé en RI par leur robustesse et leur disponibilité : un *stemmer* du français (développé par J. Savoy [SAV 99] et s'appuyant sur un ensemble de règles fixées de désuffixation), et un lemmatiseur du français (l'étiqueteur TREETAGGER[2]).

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

Le tableau 1 récapitule les résultats obtenus ; les chiffres jugés statistiquement non significatifs par un *paired t-test* [HUL 93] (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).

	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(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(500)	3.88	4.80 (+23.71%)	4.55 (+17.18%)	4.47 (+15.29%)
P(1000)	2.21	2.68 (+21.30%)	2.53 (+14.80%)	2.48 (+12.39%)
P(5000)	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(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(500)	43.09	53.83 (+24.92%)	49.31 (+14.43%)	49.35 (+14.54%)
R(1000)	48.43	59.45 (+22.74%)	55.27 (+14.12%)	55.03 (+13.62%)
R(5000)	59.32	72.20 (+21.71%)	67.22 (+13.31%)	68.20 (+14.96%)

Tableau 1. *Performances de l'extension de requête sur la collection* INIST

Notre approche obtient de très bons résultats pour chacune des mesures adoptées, tous statistiquement significatifs. On remarque notamment que l'extension de requêtes est globalement plus performante que le *stemming* ou la lemmatisation, mais aussi plus stable comme l'atteste les résultats jugés non statistiquement significatifs de ces deux techniques. Il est également intéressant de noter que l'amélioration des résultats est également distribuée sur tous les seuils de mesures (de 10 à 5000 documents). Cela signifie que l'amélioration n'est pas due à une simple réorganisation des documents en tête de liste mais bien à la découverte de documents pertinents qui n'auraient pas été ramenés par le SRI sans l'extension de requêtes.

4.2. *Influence de la taille de la requête*

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.

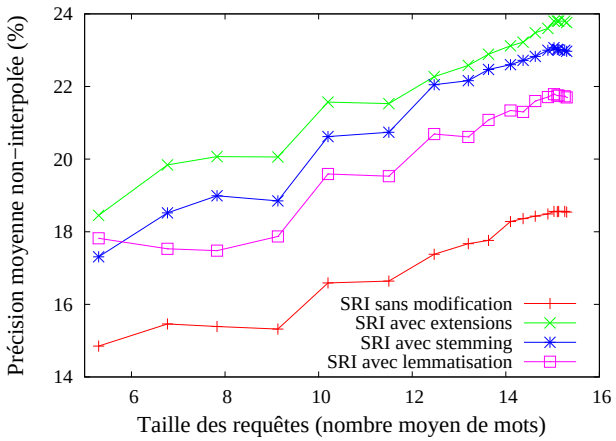

Figure 1. *Évolution de la précision selon la taille de la requête*

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

4.3. *Application à d'autres langues*

Comme nous l'avons vu, notre approche se veut portable et 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 2 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 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 variables d'une requête à une autre, ce que traduit le fait qu'ils ne soient pas jugés statistiquement significatifs.

	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(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(500)	+13.18%	+20.49%	+18.13%	+17.19%	+18.94%	+23.35%
P(1000)	+12.97%	+21.60%	+20.32%	+18.26%	+22.13%	+24.64%
R(10)	*+6.82%*	+2.90%	*+1.88%*	*+5.43%*	*-0.67%*	*-0.47%*
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(500)	+16.45%	+21.68%	+14.49%	+12.69%	+14.31%	+17.71%
R(1000)	+18.15%	+20.93%	+17.38%	+13.20%	+18.35%	+19.23%

Tableau 2. *Performances de l'extension de requête sur différentes langues*

Là encore, quelques résultats inattendus 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...). Autre résultat intéressant, l'allemand est la langue bénéficiant le plus de notre technique d'extension, sans doute par le fait que notre approche capture des cas d'agglutination fréquents dans cette langue.

4.4. *Discussion des résultats*

De ces expériences, il est difficile de tirer des conclusions allant dans le sens de celles parfois avancées dans ce domaine. D'une part, les différences de richesse morphologique tendant à opposer certaines langues ne semblent pas avoir d'influence directe dans nos expériences, contrairement à d'autres expériences [ARA 00, *inter alia*]. D'autre part, la taille des requêtes n'influe pas non plus sur nos résultats. Si les résultats sont tous positifs et constants au sein d'une collection, on constate néanmoins une variation importante de performances entre les collections 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.

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, 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 ou plus en œuvre dans la langue actuelle,

comme par exemple *composition* ↔ *exposition*, ou *pondre* ↔ *répondre*. Ensuite, certains termes polysémiques provoquent des erreurs difficiles à éviter, soulignant l'intérêt qu'il y aurait à utiliser des outils de désambiguïsation de sens. Ainsi, les mots *production* et *reproduction* trouvés comme étant liés, le sont dans *production des résultats* et *reproduction des résultats*, mais pas dans *la reproduction chez les poissons*.

5. Conclusion et perspectives

Cet article présente une technique simple permettant de détecter automatiquement des variantes morphologiques au sein de textes en s'appuyant sur la construction d'analogies dans le but d'étendre des requêtes. Les résultats obtenus par cette approche sont très satisfaisants et se comparent avantageusement aux outils supervisés testés (*stemmer* à base de règles et lemmatiseur). Enfin, la technique proposée étant non supervisée, cela nous a permis de l'appliquer à diverses langues avec là encore 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. D'un point de vue applicatif, les résultats sur les langues étudiées demandent à être vérifiés et consolidés sur d'autres collections et d'autres langues. Enfin, dans un cadre de RI translingue, l'utilisation d'une approche similaire pour la traduction de termes spécialisés est à l'étude.

6. Bibliographie

[ARA 00] ARAMPATZIS A., WEIDE T. P. V. D., KOSTER C. H. A., VAN BOMMEL P., « *Linguistically Motivated Information Retrieval* », vol. 69, p. 201-222, M. Dekker, New York, États-Unis, 2000.

[BIL 04] BILOTTI M. W., KATZ B., LIN J., « What Works Better for Question Answering : Stemming or Morphological Query Expansion ? », *Proceedings of the Information Retrieval for Question Answering (IR4QA) Workshop at SIGIR 2004*, 2004.

[CLA 05] CLAVEAU V., L'HOMME M.-C., « Structuring terminology by analogy machine learning », *Proceedings of the International conference on Terminology and Knowledge Engineering, TKE'05*, Copenhague, Danemark, 2005.

[FRE 82] FREUND G., WILLETT P., « Online Identification of Word Variants and Arbitrary Truncation Searching Using a String Similarity Measure », *Information Technology : Research and Development*, vol. 1, 1982, p. 177-187.

[FUL 98] FULLER M., ZOBEL J., « Conflation-Based Comparison of Stemming Algorithms », *Proceedings of the 3rd Australian Document Computing Symposium*, Sydney, Australie, 1998.

[GAU 00] GAUSSIER E., GREFENSTETTE G., HULL D., ROUX C., « Recherche d'information en Français et traitement automatique des langues », *Traitement automatique des langues*, vol. 41, n° 2, 2000, p. 473-493.

[HAT 01] HATHOUT N., « Analogies morpho-synonimiques. Une méthode d'acquisition automatique de liens morphologiques à partir d'un dictionnaire de synonymes », *Actes de la 8e conférence Traitement Automatique du Langage Naturel, TALN'01*, Tours, France, 2001.

[HUL 93] HULL D., « Using Statistical Testing in the Evaluation of Retrieval Experiments », *Proceedings of the 16th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'93*, Pittsburgh, États-Unis, 1993.

[HUL 96] HULL D., « Stemming Algorithms - A Case Study for Detailed Evaluation », *Journal of the American Society of Information Science*, vol. 47, n° 1, 1996, p. 70-84.

[KRO 93] KROVETZ R., « Viewing Morphology as an Inference Process », *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, États-Unis, 1993.

[LEP 03] LEPAGE Y., « De l'analogie ; rendant compte de la communication en linguistique », Thèse d'habilitation (HDR), Université de Grenoble 1, Grenoble, France, 2003.

[LOU 00] DE LOUPY C., « 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.

[LOV 68] LOVINS J. B., « Development of a Stemming Algorithm », *Mechanical Translation and Computational Linguistics*, vol. 1, 1968, p. 22-31.

[MEL 00] MEL'ČUK I., *Cours de morphologie générale*, vol. 1-5, Presses de l'Université de Montréal/CNRS Éditions, Montréal/Paris, 1993-2000.

[MOR 05] MOREAU F., SÉBILLOT P., « Contributions des techniques du traitement automatique des langues à la recherche d'information », rapport n° 1690, 2005, IRISA.

[POR 80] PORTER M., « An Algorithm for Suffix Stripping », *Program*, vol. 14, n° 3, 1980, p. 130-137.

[ROB 98] ROBERTSON S. E., WALKER S., HANCOCK-BEAULIEU M., « Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive », *Proceedings of the 7th Text Retrieval Conference, TREC-7*, 1998, p. 199-210.

[SAV 99] SAVOY J., « A stemming procedure and stopword list for general French corpora », *Journal of the American Society for Information Science*, vol. 50, n° 10, 1999.

[SAV 02] SAVOY J., « Morphologie et Recherche d'Information », Rapport technique, 2002, Institut interfacultaire d'informatique, Université de Neuchâtel.

[VIL 02] VILARES FERRO J., BARCALA M., ALONSO M. A., « Using Syntactic Dependency-Pairs Conflation to Improve Retrieval Performance in Spanish », *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics, CICLING*, Mexico, Mexique, 2002.

[XU 00] XU J., CROFT B. W., « Improving the Effectiveness of Information Retrieval with Local Context Analysis », *ACM Transactions on Information Systems*, vol. 18, n° 1, 2000, p. 79-112.