L'apport de connaissances morphologiques pour la projection de requêtes sur une terminologie normalisée

Pierre Zweigenbaum[1], Natalia Grabar[1], Stefan Darmoni[2]

[1]DIAM — SIM/DSI, Assistance Publique – Hôpitaux de Paris &
Département de Biomathématiques, Université Paris 6
{pz,ngr}@biomath.jussieu.fr

[2]Département Informatique et Réseaux, CHU de Rouen &
Laboratoire Perception, Information et Systèmes, INSA, Rouen
stefan.darmoni@chu-rouen.fr

Résumé - Abstract

L'apport de connaissances linguistiques à la recherche d'information reste un sujet de débat. Nous examinons ici l'influence de connaissances morphologiques (flexion, dérivation) sur les résultats d'une tâche spécifique de recherche d'information dans un domaine spécialisé. Cette influence est étudiée à l'aide d'une liste de requêtes réelles recueillies sur un serveur opérationnel ne disposant pas de connaissances linguistiques. Nous observons que pour cette tâche, flexion et dérivation apportent un gain modéré mais réel.

Mots Clefs - Keywords Morphologie, recherche d'information, variantes de termes, médecine, terminologie, Doc'CISMeF, MeSH.

1 Appariement entre requêtes et termes spécialisés

Une opération clé en recherche d'information est l'appariement entre une requête et un document. Comme une requête peut employer des mots proches mais non nécessairement identiques à ceux des documents, on peut espérer que la prise en compte de proximités morphologiques entre mots aide à obtenir de meilleurs résultats. Cela peut se faire par simplification préalable des requêtes et des documents (lemmatisation, « racinisation ») ou par extension de requête (ajout de formes fléchies ou dérivées aux mots de la requête). Lemmatisation et racinisation ne semblent cependant pas apporter systématiquement l'amélioration attendue ; des études complémentaires sont donc nécessaires (voir (Gaussier *et al.*, 2000) pour une synthèse et une expérimentation). Nous proposons ici une nouvelle expérimentation dans un domaine spécifique.

Son contexte est le suivant. L'annuaire des sites médicaux francophones, Doc'CISMeF (Darmoni *et al.*, 2000), indexe ces sites par des mots clés pris dans un thesaurus de référence, le

MeSH (INSERM, 2000). L'utilisateur, pour chercher ces sites, peut saisir une requête pour sélectionner les termes MeSH pertinents, qui donneront ensuite accès aux sites correspondants. La question posée est alors la suivante : l'emploi de connaissances morphologiques dans le processus d'appariement entre requête et termes cible améliore-t-il la qualité des termes trouvés ?

Cette tâche est proche de la tâche classique de recherche d'information : on part d'une requête en langue naturelle, qu'il faut apparier avec un « document » en langue naturelle. Elle possède cependant plusieurs particularités. La cible est extrêmement courte : c'est une expression comprenant de un à quelques mots. Elle n'est pas nécessairement syntaxiquement bien formée : de nombreux termes suivent un schéma du type *myocarde, infarctus* dans lequel le terme d'origine (*infarctus du myocarde*) est réduit à ses mots significatifs, le nom étant placé en tête. Une différence supplémentaire ici est que certaines requêtes sont exprimées dans une langue différente de la terminologie cible (anglais pour interroger du français). Cela vient du fait que certains utilisateurs peuvent être plus familiers avec la version originale (américaine) du MeSH qu'avec sa version française.

Cette tâche peut être rapprochée des travaux sur la variation terminologique. Cette variation a été étudiée d'un point de vue morphologique (Jacquemin & Tzoukermann, 1999), syntaxique (Jacquemin, 1997) et lexical (ou «sémantique») (Hamon *et al.*, 1998). Ici, le fait que les «termes» cible aussi bien que les requêtes ne soient pas nécessairement syntaxiquement bien formés réduit a priori l'intérêt d'approches syntaxiques.

L'appariement visé repose sur la disponibilité de ressources morphologiques. Si les travaux sur la flexion sont maintenant répandus pour le français (par exemple, (Namer, 2000)), les ressources dérivationnelles sont plus rares ; elles commencent cependant à être constituées (Gaussier, 1999; Grabar & Zweigenbaum, 2000; Daille, 1999; Hathout, 2001; Hathout *et al.*, 2001). Nous partons ici des ressources obtenues automatiquement dans (Grabar & Zweigenbaum, 2000).

Nous présentons successivement la préparation des ressources morphologiques et des requêtes, le processus d'appariement, les expériences effectuées et leurs résultats.

2 Ressources morphologiques et appariement de termes

Dans (Grabar & Zweigenbaum, 2000) et d'autres travaux connexes, nous avons extrait automatiquement de deux terminologies médicales (SNOMED et Classification internationale des maladies) des couples de mots en relation morphologique : flexion ({*abdominal*, *abdominale*}), dérivation ({*abdomen*, *abdominal*}), composition ({*adénome*, *adénofibrome*}), et des combinaisons de ces relations ({*membranes*, *membranaire*} combine la flexion {*membrane*, *membranes*} et la dérivation {*membrane*, *membranaire*}). Le principe de cette extraction automatique repose sur la disponibilité d'une terminologie possédant des relations sémantiques entre termes (synonymie, hiérarchie, etc.). Lorsque deux termes sémantiquement reliés possèdent deux mots dont la forme est proche, il y a de grandes chances que ces mots soient en relation morphologique. Par exemple, la nomenclature SNOMED indique que *sinusite, SAI*[1] est une sorte de *maladie du sinus paranasal, SAI*. On fait alors l'hypothèse que les mots {*sinus*, *sinusite*} sont en relation morphologique. On en induit aussi qu'une règle de substitution de suffixes $\epsilon | ite$ est à l'œuvre et peut s'appliquer sur d'autres couples de mots attestés du domaine. Appliquée à ces terminolo-

1. *SAI* signifie *sans autre indication*.

gies, cette méthode génère très peu de bruit (3 à 5 %) : la quasi-totalité des couples ainsi obtenus concerne des mots effectivement en relation morphologique. Nous avons également obtenu le même type de données sous forme lemmatisée (par FLEMM (Namer, 2000)) et étiquetée (avec l'aide de l'étiqueteur de Brill ; par exemple, {*phlegmon/N*, *phlegmoneux/A*}).

Comme ressources flexionnelles pour le présent travail, nous avons collecté les couples {*forme, lemme*} produits par FLEMM pour les formes fléchies des deux terminologies ci-dessus : 2906 couples correspondant à 1224 lemmes différents, soit 4125 formes en tout. Nous avons aussi utilisé les trois règles les plus fréquentes de lemmatisation : réduction d'un *-s*, *-e* ou *-es*. Pour les ressources dérivationnelles, nous avons trié et filtré 1910 couples lemmatisés et étiquetés minimaux obtenus dans (Zweigenbaum & Grabar, 2000) et contenant 2988 lemmes différents. Le filtrage concernait d'une part les quelques pourcents d'erreurs, d'autre part les combinaisons de relations morphologiques, mais aussi des couples de mots qui, bien que liés dérivationnellement, ne doivent pas être considérés comme équivalents en extension de requête dans le domaine (par exemple, {*affection/N*, *affectif/A*}). Pour séparer dérivation et composition, nous avons commencé par diviser ces couples en deux ensembles selon les deux catégories syntaxiques en présence : catégories différentes (plutôt dérivation) *vs* catégories identiques (plutôt composition savante). Nous avons ensuite ajusté manuellement cette division, et collecté 1024 couples dérivationnels (794 familles différentes, par exemple *aorte, aortique, aorto*) concernant 1759 lemmes. L'union des couples flexionnels et dérivationnels constitue 1600 familles et 5462 formes. Nous supposerons que deux formes d'un de ces couples sont substituables en recherche d'information. Notons que ces bases de connaissances lexicales n'ont pas été ajustées particulièrement au thesaurus cible de Doc'CISMeF (le MeSH) ; ce sera fait dans une expérience ultérieure. Nous avons aussi utilisé ces ressources pour déclarer les « mots vides » de l'appariement : les déterminants, prépositions, adverbes, verbes, pronoms et conjonctions.

Doc'CISMeF (http://doccismef.chu-rouen.fr) a démarré en juin 2000 ; en janvier 2001, il a reçu en moyenne plus de 1200 requêtes par jour de 400 usagers distincts. Nous avons extrait du « log » des requêtes toutes celles qui correspondaient à une recherche de termes (en supprimant celles de l'équipe CISMeF). L'expérimentation présentée ici concerne les 6469 requêtes différentes recueillies pour septembre 2000. La moitié (50 %) de ces requêtes comporte un seul mot, 29 % 2 mots, 14 % 3 mots, 5 % 4 mots et 2 % (143) comptent plus de 4 mots. Les termes cible sont ceux du MeSH français (19 971 termes, leurs synonymes et 83 qualificatifs) augmentés de 38 « métatermes » et 101 « types de ressources », soit 28 922 termes en tout (32 % comptent 1 mot, 40 % 2 mots, 18 % 3 mots, 6 % 4 mots et 3 % $>$ 4 mots).

Le processus maximal d'appariement entre requête et termes cible que nous allons tester est le suivant. Les termes cible sont pré-indexés : chaque terme est mis en minuscules, segmenté en mots (formes) et considéré comme un ensemble de mots (l'ordre n'y est plus pertinent). La forme désaccentuée de ces termes est elle aussi segmentée et indexée. Chaque requête est mise en minuscules et segmentée en mots, et les mots vides sont supprimés. Le principe de base de l'appariement est de proposer les termes qui contiennent le maximum de mots de la requête, sans tenir compte de leur position (méthode [ordre]). Plusieurs termes cible peuvent être nécessaires pour « couvrir » les différents mots d'une requête. Nous sélectionnons tour à tour (algorithme glouton) le terme cible couvrant le maximum des mots restants de la requête. L'appariement est testé sur les mots, éventuellement augmentés (extension de requête) de : (i) leur forme désaccentuée [accent] ; (ii) leurs formes fléchies [flex] ; (iii) leurs mots dérivés (et leurs formes fléchies) [deriv]. Les expériences pratiquées, comme dans (Gaussier *et al.*, 2000), consistent à ne brancher qu'une partie des étapes ([ordre] puis [accent] puis [flex] puis [deriv]) pour étudier les différences de résultats.

3 Expériences d'appariement

Nous avons effectué deux types d'observations. D'une part, l'influence quantitative de l'apport des différents paramètres d'appariement (variantes d'accentuation, flexion, dérivation) sur les résultats renvoyés : combien de réponses sont différentes lorsque telle ou telle méthode est ajoutée. D'autre part, l'effet de ces méthodes, lorsqu'elles changent la réponse obtenue, sur la correction de cette réponse. Sur les 6469 requêtes, 182 sont exactement des termes cible ; 1315 sont identiques à des termes cible une fois mises en minuscules et segmentées ; 1364 lorsqu'on les considère comme des sacs de mots [ordre] ; et 1679 une fois désaccentuées [accent] (26 %). Du fait que la terminologie cible est non accentuée, cette désaccentuation joue un rôle très important : 971 requêtes sur 6469 (15,0 %) ont des résultats différents une fois désaccentuées. Nous nous sommes focalisés ici sur l'influence de la flexion et de la dérivation.

Résultats. On observe que la flexion ([accent] → [flex]) influe sur un nombre conséquent de requêtes : 429 sur 6469 (6,6 %) voient leur résultat changer. 199 séries différentes de formes fléchies ont été employées. Il faut remarquer cependant que la quasi-totalité de l'intervention de la flexion provient des règles (403 occurrences ; 123 applications différentes du suffixe *-s*, et 48 du *-e*). Les quelques autres cas de flexion appliqués concernent des féminins : *if*|*ive* (6 occ.), *el*|*elle* (3 occ.), *ien*|*ienne* (2 occ.), *on*|*onne* (1 occ.), *blanc*|*blanche* (1 occ.), *eux*|*euse* (1 occ.). L'apport supplémentaire de la dérivation concerne un nombre plus faible, mais non négligeable, de requêtes : 128 (2,0 %). Rappelons que les mots dérivés proposés sont uniquement ceux présents dans la base lexicale constituée ci-dessus ; nous n'avons pas ici utilisé de règles pour détecter dynamiquement d'autres dérivations. On peut donc supposer que davantage de dérivations seraient potentiellement applicables. Par ailleurs, nous avons effectué les mêmes mesures sur les autres mois du log (octobre 2000 – janvier 2001), avec le même type de résultat.

Évaluation. L'évaluation manuelle a été faite par un informaticien médical et deux bibliothécaires médicaux de l'équipe CISMeF. Les requêtes comportant des fautes de frappe ont été écartées de cet examen. Pour avoir une idée de la qualité générale de l'appariement et de la difficulté de la tâche, les résultats de 58 requêtes contenant au moins 2 mots, qui n'étaient pas exactement des termes cible et ne comportaient pas de faute d'orthographe, ont été examinés. Chaque résultat a été noté de 0 (très mauvais) à 3 (très bon). La note moyenne a été de 1,7, avec un écart-type de 1,2. Les requêtes dont les résultats ont changé avec l'apport de connaissances morphologiques ont été examinées. Une évaluation a porté sur un petit échantillon de 20 requêtes sur les 429 modifiées par la flexion. Leur note moyenne a augmenté de 0,53 à 1,07. Pour la dérivation, l'évaluation a porté sur 64 des 128 requêtes concernées. La note moyenne obtenue a augmenté de 1,14 à 1,91 sur 3, une amélioration substantielle.

Le tableau 1 donne quelques exemples d'apport de connaissances dérivationnelles. Les améliorations peuvent porter aussi bien sur la précision que sur le rappel (nous séparons les réponses multiples par une barre oblique). Nous revenons sur ces résultats dans la discussion.

4 Discussion

En résumé, ces premières évaluations montrent un apport faible mais réel de la flexion (lemmatisation) et de la dérivation (racinisation) dans une tâche d'appariement de requêtes à des termes normalisés. L'une comme l'autre, lorsqu'elle agit, améliore en moyenne les réponses aux requêtes. Dans nos conditions d'expérience, la flexion agit dans 6,6 % des cas avec une

Requête	*Réponse sans dérivation*	*Réponse avec dérivation*
hematome pelvien	hematome / membre pelvien	hematome / pelvis
tumeur du glomus	glomus carotidien, tumeur	glomique, tumeur
recto colite	colite	colite / rectum
kyste du rein	rein / kyste arachnoide	rein kystique
stenose valve aorte	stenose isthmique aorte congenitale / prolapsus valve aortique	stenose aortique valvulaire
tumeur bronchique	face, tumeur / fistule bronchique	tumeur bronche
interactions entre médicaments et alimentation	interactions aliment-medicament / alimentation adolescent	interactions aliment-medicament

TAB. 1 – *Exemples de l'apport de connaissances dérivationnelles*

amélioration modérée, et la dérivation dans 2,0 % des cas avec une amélioration plus nette. Cette différence pourrait être liée aux choix différents effectués pour ces deux types de connaissances morphologiques. En effet, l'emploi de règles (non validées : *-s* et *-e*) en lemmatisation produit davantage de résultats, mais aussi davantage de bruit (par exemple, *Dublin core* → *cor triatriatum*). Pour la dérivation, nous avons uniquement employé des couples de mots validés (nous avons gardé en réserve les règles liées à ces couples de mots), pour lesquels les erreurs d'application sont plus rares (*notes personnelles* → *personne agee*).

Les résultats présentés ici sont difficiles à comparer à ceux de (Gaussier *et al.*, 2000), car les tâches abordées sont différentes. (Gaussier *et al.*, 2000) travaillent sur une tâche classique de recherche d'information : une requête étant donnée, on ordonne les documents de la base en fonction de leur pertinence espérée pour cette requête. On peut alors mesurer la précision à divers taux de rappel selon le nombre de documents que l'on conserve dans la réponse du système. Dans la tâche que nous avons examinée, les « documents » sont les termes cible ; cette tâche vise à produire un ensemble minimal de termes contrôlés « couvrant » les mots de la requête fournie. Une seule réponse est donc ramenée : en termes de recherche documentaire, elle constituerait un ensemble de documents dont chacun contient une partie de la réponse. Au-delà de son évaluation intrinsèque en tant que projection d'expressions libres sur un thesaurus contrôlé, cette tâche est utile dans une chaîne plus large de recherche d'information dont le but est de trouver les sites Web de l'annuaire CISMeF qui sont pertinents pour la requête initiale. La projection sur le thesaurus MeSH n'est alors que l'un des aspects qui doivent contribuer à identifier ces sites : à cet accès par indexation contrôlée peut s'ajouter entre autres la prise en compte d'une indexation plus traditionnelle en texte intégral.

Au-delà de ces premiers résultats, nous estimons que l'application de ces connaissances morphologiques peut encore être affinée. L'examen détaillé des situations d'erreur montre qu'une partie importante du bruit est causé par des termes complexes dont un seul mot était présent dans la requête : *tumeur bronchique* → *face, tumeur*, ou encore *grands brûlés* → *grands singes, maladies* (ou, avec flexion, *grands brûlés* → *grande bretagne* !). Nous avons donc mis en place un critère complémentaire de sélection des termes cible qui élimine ceux qui contiennent des mots non vides qui ne figurent pas dans la requête. Lorsque ce critère est activé, la flexion n'influe plus que sur 256 requêtes sur 6469 (4,0 %) ; nous en avons examiné la moitié : dans 85 % des cas, la flexion améliore les résultats, en détectant un terme supplémentaire (74 %) ou en remplaçant un terme par un autre plus précis (11 %). Les autres cas (15 %) correspondent à l'ajout d'un terme erroné. La dérivation influe pour 106 requêtes (1,6 %), que nous avons toutes examinées ; 83 % ont de meilleurs résultats, dont 8 % par précision d'un terme. Au total donc, 5,6 % des requêtes sont concernées par flexion ou dérivation.

Le filtrage supplémentaire mis en place ici rend plus clair l'apport de la morphologie : l'augmentation de rappel obtenue pour 4,7 % des requêtes s'accompagne de peu d'augmentation de bruit (0,9 % des requêtes). Cette expérience complémentaire illustre comment la spécification de la tâche et l'intervention d'autres critères de sélection des réponses peuvent modifier l'apport différentiel des connaissances morphologiques en recherche d'information.

Une démonstration de l'outil d'appariement de requêtes présenté ici est en ligne à l'adresse http://www.biomath.jussieu.fr/cismef/.

Références

DAILLE B. (1999). Identification des adjectifs relationnels en corpus. In P. AMSILI, Ed., *Actes de TALN 1999 (Traitement automatique des langues naturelles)*, p. 105–114, Cargèse: ATALA.

DARMONI S. J., LEROY J.-P., THIRION B., BAUDIC F., DOUYERE M. & PIOT J. (2000). CISMeF: a structured health resource guide. *Methods of Information in Medicine*, **39**(1), 30–35.

GAUSSIER E. (1999). Unsupervised learning of derivational morphology from inflectional lexicons. In A. KEHLER & A. STOLCKE, Eds., *ACL workshop on Unsupervised Methods in Natural Language Learning*, College Park, Md.

GAUSSIER E., GREFENSTETTE G., HULL D. & ROUX C. (2000). Recherche d'information en français et traitement automatique des langues. *Traitement automatique des langues*, **41**(2), 473–493.

GRABAR N. & ZWEIGENBAUM P. (2000). Automatic acquisition of domain-specific morphological resources from thesauri. In *Proceedings of RIAO 2000: Content-Based Multimedia Information Access*, p. 765–784, Paris, France: C.I.D.

HAMON T., NAZARENKO A. & GROS C. (1998). A step towards the detection of semantic variants of terms in technical documents. In C. BOITET, Ed., *Proceedings of the 17 th COLING*, p. 498–504, Montréal, Canada.

HATHOUT N. (2001). Analogies morpho-synonymiques. In D. MAUREL, Ed., *Actes de TALN 2001 (Traitement automatique des langues naturelles)*, Tours: ATALA et Université de Tours.

HATHOUT N., NAMER F. & DAL G. (2001). An experimental constructional database: the MorTAL project. In P. BOUCHER, Ed., *Morphology book*. Cambridge, Mass.: Cascadilla Press. *À paraître*.

INSERM (2000). *Thésaurus Biomédical Français/Anglais*. Institut National de la Santé et de la Recherche Médicale, Paris.

JACQUEMIN C. (1997). *Variation terminologique : Reconnaissance et acquisition automatique de termes et de leurs variantes en corpus*. Mémoire d'habilitation à diriger des recherches, Université de Nantes.

JACQUEMIN C. & TZOUKERMANN E. (1999). NLP for term variant extraction: A synergy of morphology, lexicon, and syntax. In T. STRZALKOWSKI, Ed., *Natural Language Processing and Information Retrieval*, p. 25–74. Boston, Mass: Kluwer.

NAMER F. (2000). FLEMM : un analyseur flexionnel du français à base de règles. *Traitement Automatique des Langues*, **41**(2), 523–547.

ZWEIGENBAUM P. & GRABAR N. (2000). Liens morphologiques et structuration de terminologie. In *IC 2000 : Ingénierie des connaissances*, p. 325–334.