

Impact des informations sémantiques sur la catégorisation automatique des documents cliniques

Natalia Grabar^{1,2}, Thierry Hamon³

¹ Centre de Recherche des Cordeliers, Université Pierre et Marie Curie - Paris6, UMR_S 872, Paris, F-75006 France ; Université Paris Descartes, UMR_S 872, Paris, F-75006 France ; INSERM, U872, Paris, F-75006 France
natalia.grabar@spim.jussieu.fr

² HEGP AP-HP, 20 rue Leblanc, Paris, F-75015

³ Laboratoire d'Informatique Médicale et Bioinformatique (EA3969)
UFR SMBH Léonard de Vinci, Université Paris 13
74, rue Marcel Cachin, 93017 Bobigny Cedex France
thierry.hamon@univ-paris13.fr

Résumé : Le codage des comptes-rendus cliniques est une tâche réalisée quotidiennement dans les hôpitaux. Nous proposons d'étudier l'influence que les informations sémantiques peuvent avoir sur ce codage. La complémentarité d'une terminologie médicale et d'une ressource de médicaments structurés est évaluée sur une collection de 1 237 comptes-rendus. Les performances du système sont améliorées avec la spécialisation des informations terminologiques.

Mots-clés : Codage médical, ressources sémantiques, catégorisation, structure hiérarchique.

Abstract : Coding of discharge summaries is performed in hospitals daily. We propose to study the impact the use of semantic information may have on this coding. The complementarity of structured medical terminology and medication resource is evaluated on a collection of 1,237 discharge summaries. Performances of the system are improved with the specialization of terminological data.

Keywords : Medical coding, semantic resources, categorization, hierarchical structure.

1 Introduction

Les hôpitaux modernes génèrent des volumes de données de plus en plus gros : comptes-rendus, prescriptions, résultats de prélèvements biologiques et anatomo-pathologiques, résumés d'imagerie, lettres aux collègues, etc. Ces don-

nées peuvent être enregistrées et stockées sous forme structurée (dates de naissance, dates de visite, noms, prénoms, adresses, ...), grâce en particulier aux systèmes d'information hospitalière et aux outils de gestion des données patient. Depuis peu, des données médicales (diagnostics, observations, ...) peuvent aussi être enregistrées sous forme structurée grâce à l'utilisation de questionnaires spécifiques (Chen *et al.*, 2008). Mais la plupart des données médicales et biologiques, pertinentes pour une tâche, restent enfouies au milieu des données non structurées. L'accès à ces données est un enjeu toujours réel et requiert des méthodes et algorithmes appropriés. Dans ce travail, nous nous sommes intéressés à l'extraction d'information à partir des documents cliniques et plus particulièrement à la détection des pathologies et co-morbidités. Cette tâche est représentative des types de tâches que les cliniciens et chercheurs effectuent en routine dans les hôpitaux dans de différents contextes, comme : codage médico-économique (codage des dossiers patient pour rendre compte de l'activité d'un hôpital), recherche clinique (constitution de cohortes de patients ayant des caractéristiques données), médecine translationnelle (création d'un lien plus direct entre la recherche thérapeutique et les soins des patients, afin de les faire bénéficier des traitements plus performants et appropriés (Rader & Daugherty, 2008)).

L'utilisation des connaissances spécifiques aux domaines, souvent proposées par les terminologies, est alors indispensable. La projection et reconnaissance de termes devient ainsi une tâche importante et il a été montré que l'utilisation de synonymes ou l'*explosion du concept* améliorent les performances de systèmes automatiques (Jacquemin & Tzoukermann, 1999; Grau *et al.*, 2005). Mais les terminologies proposent d'autres informations sémantiques qui restent relativement peu exploitées. Nous sommes tout particulièrement intéressés par l'exploitation de la structure hiérarchique des terminologies et de la spécificité des termes. De telles données commencent à être exploitées dans la langue générale (Baziz *et al.*, 2007). Quant au domaine biomédical, plusieurs travaux ont été réalisés grâce à l'accessibilité de terminologies structurées. Ainsi, (Masys *et al.*, 2001) ont exploité la structure hiérarchique du MeSH pour interpréter les résultats sur les puces ADN. Les résultats montrent que l'utilisation des niveaux hiérarchiques supérieurs apporte un biais : ils induisent une sur-représentation à cause des termes qu'ils subsument. Il apparaît également que l'exploitation des informations hiérarchiques est particulièrement bénéfique lorsqu'elles sont combinées à des informations linguistiques : collocations, dépendances syntaxiques, lemmes (Stevenson *et al.*, 2008) ; ou des données biologiques (Savova *et al.*, 2008). Par ailleurs, l'utilisation de différents types de termes MeSH (termes spécifiques (*Main Headings*) ou complémentaires) influence les performances des systèmes de manière non homogène (Funk *et al.*, 1983; Névéal *et al.*, 2007).

Dans notre travail, nous proposons d'appliquer les méthodes de catégorisation automatique de documents cliniques et, plus particulièrement, d'étudier l'impact de la spécificité des termes sur la détection de diagnostics dans des documents cliniques. Nous travaillons avec les documents fournis par le *I2B2 2008 challenge*¹, orienté sur la détection des cas d'obésité et de ses co-morbidités.

¹<http://www.i2b2.org/NLP/Obesity>

2 Matériel

Comptes-rendus et données de référence

Les comptes-rendus fournis par le *I2B2 challenge* sont écrits en anglais. Ils ont été anonymisés (Sibanda & Uzuner, 2006). Ces documents (n=1 237) se composent de deux ensembles : documents d'entraînement (n=730) et documents de test (n=507). Ces données ont été annotées : les pathologies recherchées sont indiquées comme *Y* présentes, *N* absentes ou *Q* possibles (*questionnables*). Le statut *Q* correspond à des situations où, suite à la lecture d'un compte-rendu, une pathologie est suspectée ou bien elle paraît ne pas être acquise définitivement. Ces annotations de référence ont été effectuées manuellement par deux experts médicaux. En cas de divergence inter-expert, un consensus a été établi. La distribution des annotations entre les catégories *Y*, *N* et *Q*, toute pathologie confondue, n'est pas homogène. Par exemple, dans l'ensemble d'entraînement (10 655 annotations), nous observons 69 % (n=7,362) d'annotations *N*, 31 % (n=3,267) d'annotations *Y* et seulement 0.2 % (n=26) d'annotations *Q*.

Terminologies

Pathologies. Nous utilisons la liste des 16 pathologies du *I2B2 challenge* relatives à l'obésité et ses 15 co-morbidités les plus fréquentes : asthma (asthme), CAD (maladie de l'artère coronaire), CHF (défaillance cardiaque congestive), depression (dépression), GERD (reflux gastro-oesophagien), gallstones (lithiase biliaire), gout (goutte), hypercholesterolemia (hypercholestérolémie), hypertension (hypertension), hypertriglyceridemia (hypertriglycéridémie), OA (ostéoarthrite), OSA (apnée obstructive du sommeil), PVD (angiopathie périphérique) et venous insufficiency (insuffisance veineuse).

Termes de la Snomed Internationale. Nous utilisons 108 531 termes provenant de six axes de la Snomed Internationale (SNMI) (Côté *et al.*, 1997) : Agents et activités physiques (*A*), Diagnostics (*D*), Fonction (*F*), Morphologie (*M*), Procédures (*P*) et Topographie (*T*). Nous n'utilisons pas les autres axes sémantiques (*i.e.*, Modificateurs (*G*), Métiers (*J*), Social (*S*)), dont les termes nous paraissent moins pertinents pour le codage médical. Les termes utilisés sont enrichis avec leurs synonymes d'UMLS (NLM, 2008). De plus, chaque terme SNMI est associé à son type sémantique. Par exemple, *decreased thickness* et *obstructive sleep apnea* appartiennent respectivement aux axes *M* et *D* et ont les identifiants SNMI M-02630 et D9-74310. Nous utilisons cinq niveaux hiérarchiques des termes SNMI, qui peuvent être détectés grâce à la convention d'encodage de la hiérarchie dans les identifiants SNMI. Le principe général est le suivant : plus un terme est générique plus il a de 0 à la fin de son identifiant. Par exemple, M-02000 est plus générique que M-02600, qui est plus générique que M-02630. Le niveau hiérarchique le plus haut correspond aux axes sémantiques SNMI (*i.e.*, *M*, *D*, *T*). Ainsi, pour *decreased thickness* M-02630 les niveaux hiérarchiques sont : M, M-02000, M-02600, M-02630 et M-02630 ; pour *obstructive sleep apnea* ils sont D, D9-74000, D9-74300, D9-74310 et D9-74310. Dans ces deux exemples, les termes les plus spécifiques se positionnent au niveau TSS4, le niveau TSS5

ne peut donc pas être spécifié davantage.

Noms des médicaments. Nous exploitons une liste de 5 716 DCI (dénomination commune internationale) des médicaments. Ils proviennent de la ressource Thériaque². Comme pour les termes SNMI, chaque DCI est associé à son information sémantique : les classes thérapeutiques définies par Thériaque. Par exemple, *Allopurinol* est un médicament antigoutte.

3 Méthode

Pré-traitement et indexation des documents

L'objectif du pré-traitement des comptes-rendus est de les indexer avec les ressources terminologiques. Nous utilisons la plate-forme Ogmios³ pour : (1) Segmentation des comptes-rendus en mots et phrases ; (2) Étiquetage morpho-syntaxique et lemmatisation (Schmid, 1994) pour la normalisation syntaxique (réduction de pluriels et de formes fléchies) ; (3) Étiquetage terminologique avec la projection des entités des ressources utilisées ; (4) Détection des variantes de termes (Jacquemin & Tzoukermann, 1999) : {*hernia repair*, *repair of ventral hernias*}, {*weight gain*, *gaining weight*} ; (5) Détection de la négation (Chapman *et al.*, 2001). Les étapes 2 à 4 permettent de détecter les groupes nominaux, de les normaliser et de les proposer pour l'indexation des documents. En résultat, à chaque compte-rendu est associé un ensemble d'entités : pathologies du challenge, termes SNMI et DCI.

Catégorisation des documents

Définition des descripteurs. Deux types de descripteurs sont utilisés, ce qui fournit six ensembles de descripteurs :

T : termes SNMI, pathologies du challenge, DCI (n=4 165)

TSS : niveaux hiérarchiques des descripteurs. Nous utilisons cinq niveaux hiérarchiques pour les termes SNMI et un niveau hiérarchique pour les DCI.

Nous obtenons cinq ensembles de descripteurs TSS1, TSS2, TSS3, TSS4 et TSS5, où TSS1 est le niveau le plus générique et TSS5 est le niveau le plus spécifique. Au niveau le plus générique, nous avons 440 descripteurs.

Les valeurs des descripteurs sont : (1) statut des descripteurs (*y* contexte positif, *n* contexte négatif, 0 absents) ; ou (2) fréquences des descripteurs pour les contextes positifs et les contextes négatifs séparément.

Algorithmes d'apprentissage. Après avoir testé plusieurs algorithmes d'apprentissage supervisé proposés par la plate-forme Weka (Witten & Frank, 2005), nous avons choisi d'utiliser l'algorithme à base d'arbres de décision REPTree (variante de C4.5 (Quinlan, 1993)). Pour chaque pathologie, nous avons construit cinq modèles (ou arbres de décision) qui contrastent un type d'annotations avec les autres.

²<http://www.theriaque.org>

³<http://search.cpan.org/~thamon/Alvis-NLPPlatform/>

Évaluation

L'évaluation des résultats, à savoir si chaque document est assigné aux bonnes catégories, est effectuée par les organisateurs du challenge selon les annotations de référence de l'ensemble de test. Les mesures d'évaluation standard sont utilisées (R rappel, P précision) et la moyenne harmonique F (F-mesure). Ces mesures sont appliquées en version micro (la moyenne est calculée au niveau des documents) et macro (la moyenne est calculée au niveau des catégories Y , Q , N et U) (Sebastiani & Ricerche, 2002). La mesure principale est macro F-mesure globale, qui correspond à la moyenne des macro F-mesures de chaque pathologie. Pour chaque couple de résultats, nous calculons également la significativité statistique des résultats afin de mieux observer l'apport des informations sémantiques. La distribution de données nous étant inconnue, nous utilisons le test de significativité par répartition au hasard (Chinchor, 1992) utilisé dans les compétitions MUC. Ainsi, après avoir calculé la différence de résultats entre les deux systèmes à comparer, n couples de pseudo-systèmes (fixé à 1000) ont été générés aléatoirement à partir des deux systèmes existants. On compte alors le nombre c de fois que la différence entre les deux pseudo-systèmes d'un même couple est supérieure à la différence des deux systèmes existants. Si le ratio entre le nombre de c et n est inférieur au seuil de rejet (fixé à 0,05) de l'hypothèse nulle, les deux systèmes sont alors significativement différents.

4 Résultats et discussion

Les 507 comptes-rendus de l'ensemble de test ont été traités par la plate-forme Ogmios en 53 min 19 sec (6.3 sec par documents).

Résultats par pathologie et par catégorie

	MiP	MaP	MiR	MaR	MiF	MaF
Notre système	0.92	0.62	0.91	0.61	0.91	0.61
Moyenne du challenge	0.91	0.78	0.90	0.60	0.90	0.60
Écart type	0.09	0.17	0.09	0.06	0.09	0.06

TAB. 1 – Résultats pour toutes les pathologies.

La table 1 indique les performances de notre système, la moyenne des résultats du challenge et leur écart type (ou dispersion autour de la moyenne). La macro F-mesure de notre système, de même que toutes les mesures sauf la macro précision, ont des valeurs supérieures aux moyennes correspondantes. Notons que les macro F-mesures du challenge s'étendent entre 0.67 et 0.32 : notre système est très proche de la meilleure valeur.

La figure 1 présente les performances de notre système par pathologie, en termes de macro précision et de macro rappel. Dans cette représentation, meilleurs

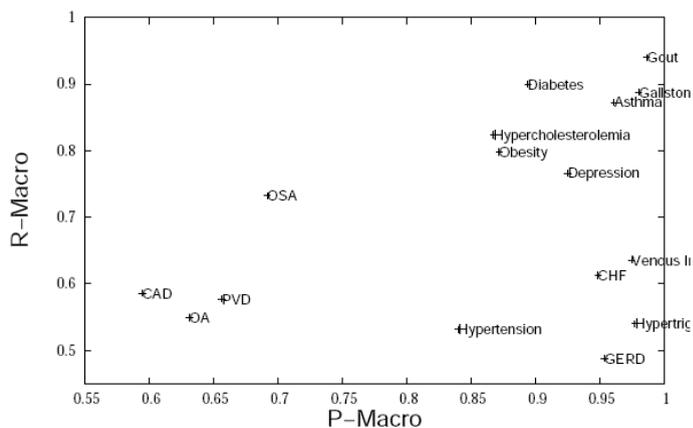


FIG. 1 – Macro précision et macro rappel par pathologie.

sont les résultats pour une pathologie donnée, plus cette pathologie est proche du coin supérieur droit du graphique. À l'analyse de cette figure, nous pouvons distinguer trois zones :

1. zone proche du coin supérieur droit : la qualité de notre système est bonne avec les annotations (pathologies et catégories). La macro précision et le macro rappel sont supérieurs à 0.80. Cela concerne les pathologies : asthme, dépression, diabète, lithiase biliaire, goutte, hypercholestérolémie et obésité.
2. zone vers le coin inférieur droit : la précision reste élevée (> 80 %) mais le rappel diminue (entre 50 % et 70 %). Dans cette zone, nous trouvons : défaillance cardiaque congestive, reflux gastro-oesophagien, hypertension, hypertriglycéridémie et insuffisance veineuse.
3. zone vers le coin inférieur gauche : les valeurs de la précision et du rappel sont entre 50 % et 70 %. Quatre pathologies se trouvent dans cette zone : maladie de l'artère coronaire, ostéoarthrite, apnée obstructive du sommeil et angiopathie périphérique. Elles correspondent aux cas les plus difficiles à catégoriser pour notre système.

L'utilisation de termes et de leurs synonymes pour l'indexation des documents nous place dans le cadre d'une approche *basée sur les concepts*. Si cela permet de restreindre le champs sémantique des descripteurs, comparé à une approche à base de *sacs de mots*, et de le rendre ainsi plus précis, nous nous trouvons limités par la variabilité terminologique. Cette limitation peut résulter de la couverture partielle des terminologie (Hersh *et al.*, 1994) ou de l'incapacité des outils automatiques à reconnaître dans les documents les termes des terminologies utilisées. Dans notre travail, nous avons mis l'accent sur la reconnaissance automatique des variantes de termes. Néanmoins, pour une pathologie comme OSA, nous ob-

tenons un ensemble de descripteurs décisionnels assez pauvre (*obstructive sleep apnea, sleep apnea, apnea, OSA, pain*). De plus, la variation terminologique ne suit pas toujours des règles précises et reproductibles, comme par exemple dans les cas de certaines abréviations “cliniques” (*HTN* pour *hypertension*). Mais nous pensons qu’il est possible d’apporter des améliorations à cette fonctionnalité de la plate-forme grâce à l’intégration d’autres ressources et outils.

		Notre système			
		Y	N	Q	NR
Référence	Y	1846	388	35	0
	N	162	4816	16	0
	Q	3	10	1	0
	NR	224	467	6	0

TAB. 2 – Matrice de confusion par catégorie (Y , N , Q).

La table 2 présente la matrice de confusion par catégorie (Y , N et Q). Les annotations de référence sont indiquées en ligne, les annotations effectuées par notre système sont indiquées en colonne. Les nombres en gras correspondent aux annotations correctes : 1 846 pour Y , 4 816 pour N et seulement une pour Q . Les autres nombres, lus en colonne, s’ils sont supérieurs à zéro, indiquent les erreurs de notre système. Par exemple, 388 documents reçoivent une annotation N et 35 une annotation Q , alors que leur annotation de référence est Y . De la même manière, 162 annotations Y et 16 Q sont en réalité des annotations N . Cela nous donne une précision et un rappel entre 83 et 96 % pour les catégories Y et N , tandis que pour la catégorie Q , ils sont inférieurs à 10 %. Rappelons que la catégorie Q concerne l’annotation des pathologies qui ne sont pas acquises de manière certaine mais probable. Le marquage de la certitude (Hyland, 1995; Fahnstock, 1999) passe souvent par l’utilisation des marqueurs comme *certainly, likely, might* ou *seem* dans la littérature scientifique. Un examen des comptes-rendus cliniques ayant des annotations Q montre que la certitude n’est pas exprimée de la même manière. Par exemple, l’annotation Q des cas de diabète semble être justifiée par des phrases en relation avec la nutrition : (1) *She also was given nutrition consult who started the calorie counts.* (2) *Nutrition recommended that she may be need long-term nutritional therapy if she does not meet her nutritional goal.* (3) *DIET : House / Low chol/low sat. fat* Nous pouvons même avoir des phrases de type *nondiabetic ... in 1986* sans autre indication sur le cas de diabète dans le compte-rendu. Pour les cas d’obésité, nous pouvons avoir des phrases comme (1) *claiming that he has been eating like a horse* ou (2) *The patient denies weight gain or nausea.* Comme le montrent ces exemples, les indicateurs du statut Q sont vagues et pauvres dans les comptes-rendus. Leur détection peut difficilement reposer sur des marqueurs de certitude utilisés traditionnellement mais nécessiterait plutôt un raisonnement médical fin.

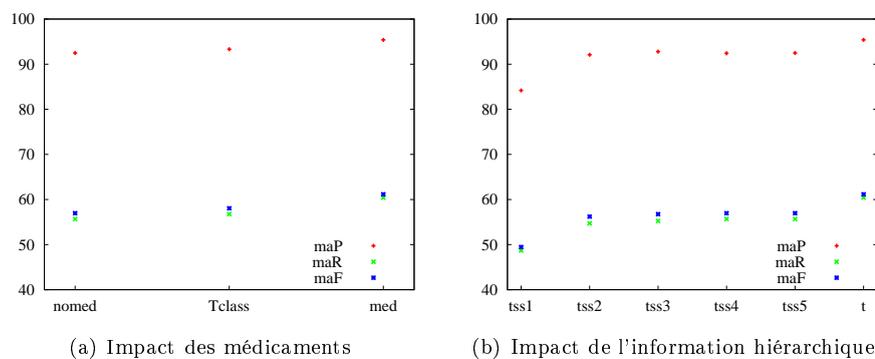


FIG. 2 – Impact des informations sémantiques.

Prise en compte des médicaments

Les médicaments (leurs prescriptions, prises, contre-indications, arrêt, etc.) font partie intégrante de tout dossier médical. Les noms de médicaments sont des nomenclatures ou listes, qui peuvent bénéficier d'une organisation hiérarchique. Dans la ressource Thériaque, que nous utilisons, les médicaments sont organisés en classes thérapeutiques : médicaments anti-inflammatoires, antigoutteux, digestifs, laxatifs, antiseptiques.

Afin d'étudier l'impact des informations sur les médicaments, nous avons comparé les résultats obtenus avec trois expériences (fig. 2(a)) : les descripteurs de pathologies du niveau TSS5 sans les médicaments (*nomed*), avec les classes thérapeutiques (*Tclass*) et avec les noms de médicaments (*med*). Nous indiquons les valeurs de macro précision, macro rappel et macro F-mesure. Avec l'injection progressive des informations sur les médicaments, nous pouvons observer une amélioration des performances. L'utilisation des classes thérapeutiques et des noms de médicaments apporte respectivement un gain de 1,5 et 4,19 à la F-mesure. Ces différences sont statistiquement significatives. L'évolution de la précision et du rappel suit le même schéma. Notons néanmoins que la ressource Thériaque n'est pas exhaustive et que certains médicaments, comme *coumadin*, *torsemide* ou *zaroxolym*, n'y figurent pas. Nous pensons que l'impact des médicaments pourrait être plus important si une liste plus complète, comme Rxnorm (RxNorm, 2009), était utilisée.

Prise en compte des informations hiérarchiques

Afin d'étudier l'influence de l'informativité des termes, nous observons les annotations obtenues à différents niveaux hiérarchiques des termes d'indexation (fig. 2(b)). Pour ceci, nous comparons plusieurs ensembles d'annotations obtenues avec l'algorithme REPTree : cinq niveaux hiérarchiques de la SNMI (TSS1 à TSS5) avec les classes thérapeutiques des médicaments, et T (le niveau le

plus spécifique de SNMI avec les noms de médicaments). La différence la plus importante apparaît entre les niveaux TSS1 et TSS2, où le niveau TSS1 est le plus générique. Une autre amélioration est observée entre les niveaux TSS5 et T. Dans ces deux cas, la différence est statistiquement significative. Dans tous les cas, la spécification de l'annotation est profitable aux performances du système. La F-mesure passe par exemple de 49,46 avec TSS1 à 61,15 avec T. L'évolution de la précision et du rappel est similaire.

Ces résultats confirment les observations de (Stevenson *et al.*, 2008) et (Savova *et al.*, 2008) selon lesquelles les informations linguistiques contribuent fortement à l'amélioration des performances d'un système basé sur l'exploitation des informations hiérarchiques d'une terminologie. Nous observons également que l'ensemble de ces informations hiérarchiques ne contribuent pas de la même manière : des informations trop générales dégradent les performances. Il est donc important de bien sélectionner les niveaux hiérarchiques à utiliser lors de la catégorisation des documents.

5 Conclusion et perspectives

Nous avons présenté une expérience de catégorisation automatique des documents cliniques en anglais, évaluée selon les données de référence produites par le challenge i2b2 en 2008. Au travers du traitement de ces données, nous testons plus particulièrement l'exploitation des informations sémantiques et linguistiques (variation des termes, médicaments, niveaux hiérarchiques des termes) et observons que, de manière générale, la spécificité de ces informations et l'utilisation de ressources complémentaires montrent une influence bénéfique. Les performances de notre système sont légèrement supérieures à la moyenne obtenue lors du challenge. Nous envisageons néanmoins d'améliorer ces performances et pour ceci nous visons une meilleure détection des variantes de termes, une sélection de descripteurs et l'exploitation de la structure de comptes-rendus. Par ailleurs, l'impact d'autres facteurs, comme la négation ou le vote des algorithmes, peut aussi avoir une incidence sur les performances du système.

Références

- BAZIZ M., BOUGHANEM M., PASI G. & PRADE H. (2007). An information retrieval driven by ontology : from query to document expansion. In *Proceedings of RIAO 2007*, Pittsburgh, USA.
- CHAPMAN W., BRIDEWELL W., HANBURY P., COOPER G. & BUCHANAN B. (2001). Evaluation of negation phrases in narrative clinical reports. In *Annual Symposium of the American Medical Informatics Association (AMIA)*, Washington.
- CHEN R., GARDE S., BEALE T., NYSTRÖM M., KARLSSON D., KLEIN G. O. & ÄHLFELDT H. (2008). An archetype-based testing framework. In *MIE 2008, IOS Press*, p. 401-6.
- CHINCHOR N. (1992). The statistical significance of the MUC-4 results. In *MUC*, p. 30-50.

- CÔTÉ R. A., BROCHU L. & CABANA L. (1997). *SNOMED Internationale – Répertoire d'anatomie pathologique*. Secrétariat francophone international de nomenclature médicale, Sherbrooke, Québec.
- FAHNESTOCK J. (1999). *Rhetorical figures in science*. Oxford University Press.
- FUNK M. E., REID C. A. & MCGOOGAN L. S. (1983). Indexing consistency in medline. *Bull Med Libr Assoc*, **71**(2), 176–183.
- GRAU B., ROBBA I., VILNAT A., KATEB F. E. & PAROUBEK P. (2005). De l'importance des synonymes pour la sélection de passages en question-réponse. In *CORIA*, p. 71–84, Grenoble, France.
- HERSH W. R., HICKAM D. H., BRIAN HAYNES & MCKIBBON K. A. (1994). A performance and failure analysis of SAPHIRE with Medline test collection. *JAMIA*, **1**(1), 51–60.
- HYLAND K. (1995). The author in the text : Hedging scientific writing. *Hong Kong papers in linguistics and language teaching*, **18**, 33–42.
- 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*, chapter 2, p. 25–74. Boston, MA : Kluwer.
- MASYS D. R., WELSH J. B., FINK J. L., GRIBSKIOV M., KLCANSKY I. & CORBEIL J. (2001). Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics*, **17**(4), 319–326.
- NÉVÉOL A., MORK J. G. & ARONSON A. R. (2007). Automatic indexing of specialized documents : using generic vs. domain specific document representations. In *BIONLP 2007*, p. 183–190.
- NLM (2008). *UMLS Knowledge Sources Manual*. National Library of Medicine, Bethesda, Maryland. www.nlm.nih.gov/research/umls/.
- QUINLAN J. (1993). *C4.5 Programs for Machine Learning*. San Mateo, CA : Morgan Kaufmann.
- RADER D. & DAUGHERTY A. (2008). Translating molecular discoveries into new therapies for atherosclerosis. *Nature*, **451**(7181), 904–13.
- RxNorm (2009). *RxNorm, a standardized nomenclature for clinical drugs*. Rapport interne, National Library of Medicine, Bethesda, Maryland. Disponible à www.nlm.nih.gov/research/umls/rxnorm/docs/index.html.
- SAVOVA G., CLARK C., ZHENG J., COHEN K. B., MURPHY S., WELLNER B., HARRIS D., LAZO M., ABERDEEN J., HU Q., CHUTE C. & HIRSCHMAN L. (2008). The Mayo/MITRE System for Discovery of Obesity and Its Comorbidities. In *The Second i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*.
- SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, p. 44–49, Manchester, UK.
- SEBASTIANI F. & RICERCHÉ C. N. D. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, **34**, 1–47.
- SIBANDA T. & UZUNER O. (2006). Role of local context in de-identification of ungrammatical, fragmented text. In *Proceedings of the North American Chapter of Association for Computational Linguistics/Human Language Technology (NAACL-HLT 2006)*, New York, USA.
- STEVENSON M., GUO Y., GAIZAUSKAS R. & MARTINEZ D. (2008). Knowledge sources for word sense disambiguation of biomedical text. In *BioNLP 2008*, p. 80–87, Columbus, Ohio, USA.
- WITTEN I. & FRANK E. (2005). *Data mining : Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco.