Aide à l'enrichissement d'un référentiel terminologique : propositions et expérimentations

Thibault Mondary[1] Adeline Nazarenko[1]
Haïfa Zargayouna[1] Sabine Barreaux[2]
(1) Université Paris 13, Sorbonne Paris Cité, LIPN (UMR 7030), F-93430 Villetaneuse, France
(2) INIST-CNRS, Vandœuvre-lès-Nancy, France
(1) prenom.nom@lipn.univ-paris13.fr, (2) prenom.nom@inist.fr

RÉSUMÉ

En s'appuyant sur une expérience d'enrichissement terminologique, cet article montre comment assister le travail d'acquisition terminologique et surmonter concrètement les deux difficultés qu'il présente : la masse de candidats-termes à considérer et la subjectivité des jugements terminologiques qui varient notamment en fonction du type de terminologie à produire. Nous proposons des stratégies simples pour filtrer *a priori* une partie du bruit des résultats des extracteurs et rendre ainsi la validation praticable pour des terminologues et nous démontrons leur efficacité sur un échantillon de candidats-termes proposés à la validation de deux spécialistes du domaine. Nous montrons également qu'en appliquant à une campagne de validation terminologique les mêmes principes méthodologiques que pour une campagne d'annotation, on peut contrôler la qualité des jugements de validation posés et de la terminologie qui en résulte.

ABSTRACT

Help enrich a terminological repository : proposals and experiments

Based on an experience of terminological enrichment, this paper shows how to support the work of terminological acquisition and overcome practical difficulties it presents, *i.e.* the mass of candidate terms to consider and the subjectivity of terminological judgments which depends on the type of terminology to produce. We propose simple strategies to filter *a priori* part of the noise from the results of term extractors so as to make the validation practicable for terminologists. We demonstrate their effectiveness on a sample of candidate terms proposed for the validation of two experts. We also show that by applying to term validation campaigns the methodological principles that have been proposed for corpus annotation campaigns, we can control the quality of validation judgments and of the resulting terminologies.

MOTS-CLÉS : Acquisition terminologique, validation de candidats-termes, filtrage de termes, distance terminologique, vote, accord inter-juges.

KEYWORDS: Terminology acquisition, term candidate validation, term filtering, terminological distance, vote, inter-judge agreement.

1 Introduction

Les ressources terminologiques, qu'elles soient monolingues, bilingues ou autres, sont utilisées dans de nombreux outils de gestion de contenus spécialisés mais leur élaboration présente souvent un coût rédhibitoire. La mise à disposition de ressources [1] ne résout que partiellement le problème car l'évolution des domaines et des besoins applicatifs rend nécessaires de fréquentes mises à jour.

Des outils d'extraction terminologiques ont été développés depuis une vingtaine d'années (Jacquemin et Bourigault, 2003) pour automatiser les processus d'acquisition terminologique mais on sait que les extracteurs de termes ne peuvent fournir au mieux que des « candidats termes », que des mots ou groupes de mots qui, sur la base de propriétés syntaxiques, lexicales et statistiques, semblent avoir un comportement terminologique, c'est-à-dire avoir un sens précis et relativement stable au sein d'un domaine de spécialité.

L'acquisition d'une terminologie pour un domaine particulier à l'aide d'outils d'analyse terminologique se heurte en fait à une double difficulté. La première concerne le filtrage et le retraitement des sorties d'analyseurs qui demandent à être validées par un terminologue si on vise une terminologie de qualité et consultable [2]. Ce travail de validation peut s'avérer très fastidieux quand on utilise de gros corpus d'acquisition et que les extracteurs utilisés sont prolixes. La seconde difficulté est liée à la diversité des styles terminologiques : il existe des terminologies de taille très variable, même pour un même domaine ; la granularité de la description terminologique varie ; certaines terminologies recensent toutes les variantes des termes alors que d'autres ne listent que les termes canoniques ou « recommandés » ; dans une perspective d'annotation sémantique, on privilégie les termes longs, alors qu'on préférera des termes plus courts pour les tâches d'indexation. Le choix d'un style de terminologie n'est généralement pas guidé par les outils d'extraction terminologique mais il faut néanmoins en tenir compte dans le travail de validation.

En s'appuyant sur une expérience d'enrichissement terminologique menée en collaboration entre l'INIST et le LIPN [3], cet article montre comment on peut concrètement surmonter ces deux difficultés et assister le travail d'acquisition terminologique. La section 2 présente le contexte dans lequel cette expérience a été menée puis nous montrons comment on peut filtrer *a priori* une partie du bruit des résultats des extracteurs pour rendre la tâche de validation accessible à des spécialistes du domaine (section 3) tout en contrôlant la qualité, ou du moins l'homogénéité, de ce travail (section 4).

2 Contexte expérimental

La question de l'évolution des référentiels d'indexation est une question importante pour tout organisme qui gère et maintient de tels référentiels. C'est en particulier le cas de l'INIST. A partir d'un thésaurus de pharmacologie utilisé comme référentiel d'indexation, deux questions se sont

1. Par exemple par l'Office québécois de la langue française (http ://gdt.oqlf.gouv.qc.ca/) ou la Délégation générale à la langue française et aux langues de France (http ://www.culture.fr/Ressources/FranceTerme).

2. Les sorties des extracteurs peuvent parfois être utilisées telles que quand elles sont directement intégrées dans des systèmes qui sont robustes au bruit, par exemple certaines application de classification de documents.

3. Ce travail s'inscrit dans le prolongement des campagnes d'évaluation des outils d'extraction terminologique menées dans le cadre du programme Quaero (projets CTC et Corpus). Il a été en partie financé par ce programme.

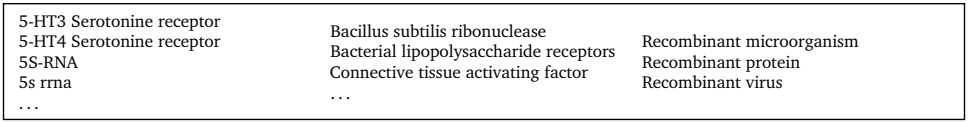

FIGURE 1 – Extrait du référentiel terminologique

posées. Est-il possible d'assister la mise à jour de ce référentiel qui se faisait jusque là de manière purement manuelle ? Est-il possible de construire à partir de ce thésaurus une terminologie adaptée à des tâches d'annotation sémantique ? Avec ces objectifs en tête, nous avons cherché à définir un protocole d'enrichissement terminologique qui tire le meilleur parti de l'expertise des terminologues et assure un travail de qualité.

Le référentiel terminologique Le référentiel est un thésaurus construit par l'INIST à des fins d'indexation de la partie pharmacologique de la base de données bibliographiques PASCAL [4]. Il contient 76 466 termes en anglais avec certaines variations et certaines relations hiérarchiques, et est accessible *via* TermSciences [5], le portail terminologique multidisciplinaire mis en place par l'INIST. Nous l'utilisons ici comme simple terminologie, sans tenir compte des relations terminologiques qu'il comporte. Un extrait est présenté sur la figure 1. Ce référentiel d'indexation privilégie les termes généraux du domaine de la pharmacologie au détriment des termes très spécifiques.

Les corpus d'acquisition Le processus d'extraction de termes repose sur l'existence de corpus d'acquisition. Dans le cadre de cette expérience, deux corpus anglais ont été utilisés. Le premier (corpus CR) est constitué de résumés d'articles de pharmacologie de la base PASCAL, le genre de textes couramment utilisé par l'INIST pour l'indexation des articles scientifiques. Il comporte 1 500 000 mots. Le second corpus (CB) porte aussi sur la pharmacologie mais il est composé de textes différents. Il s'agit de brevets européens qu'il est prévu d'annoter sémantiquement dans le cadre du programme Quaero. Il comporte 2 500 000 mots.

Les extracteurs de termes Les extracteurs de termes utilisent différentes stratégies pour extraire des candidats-termes. Certains comme YaTeA (Aubin et Hamon, 2006) ou Acabit (Daille, 2003) utilisent des patrons linguistiques, tandis que d'autres comme Termostat (Drouin, 2006) reposent sur l'analyse des contrastes entre un corpus de domaine général et un corpus de spécialité. Quasiment tous utilisent des filtrages statistiques avec des seuils plus ou moins tolérants afin de filtrer le bruit en fonction de l'objectif visé par l'extracteur (par exemple un petit nombre de candidats-termes potentiellement représentatifs, ou alors une couverture maximale). Nous avons observé une grande hétérogénéité dans le nombre de termes extraits sur un même corpus, certains extracteurs produisant 200 fois plus de termes que d'autres.

Dans cette expérience, nous avons utilisé les sorties des extracteurs testés lors de la campagne Quæro (Mondary *et al.*, 2012) [6]. Les différentes stratégies d'extraction sont représentées. Dans

4. http://inist.fr/spip.php?article170
5. http://www.termsciences.fr
6. Notamment Acabit, Termostat et YaTeA, ainsi que des prototypes de recherche des partenaires Quæro.

l'ensemble, les extracteurs sont verbeux. Les corpus de résumés (CR) et de brevets (CB) ont permis respectivement d'extraire 321 124 et 303 648 candidats-termes. L'union des sorties des extracteurs sur les deux corpus donne un total de 570 608 candidats-termes différents. Certains de ces candidats-termes existaient déjà dans le référentiel de l'INIST mais un nombre significatif de nouveaux termes ont été proposés : 298 593 et 271 472 candidats-termes resp. pour CR et CB.

L'interface de validation L'objectif étant de valider les nouveaux termes extraits, une interface de validation a été fournie aux experts de l'INIST. C'est une application web, qui est disponible sur Sourceforge. ValiTerms [7] permet aux terminologues de visualiser les occurrences des candidats-termes à valider dans leur contexte (les phrases du corpus) et offre la possibilité de choisir pour chaque terme s'il est correct, incorrect ou douteux [8]. Une zone de texte en face de chaque terme permet éventuellement d'indiquer la forme correcte attendue.

3 Filtrer *a priori* une partie du bruit

Il n'est pas raisonnable de demander à des experts de valider plusieurs centaines de milliers de candidats-termes. Nous devons trouver des stratégies pour proposer à l'expert les candidats-termes les plus à même de l'intéresser.

3.1 Deux hypothèses à valider

Filtrer par le vote des systèmes Dans la mesure où nous disposions des sorties de plusieurs extracteurs, nous avons proposé une première stratégie de filtrage consistant à donner en priorité à valider aux terminologues les termes retrouvés par plus de systèmes. C'est une technique de vote classique (Choi, 1999). L'intuition est que les candidats-termes retrouvés par plusieurs systèmes ont plus de chance d'être représentatifs que les candidats-termes retournés par un seul extracteur, même si un biais de cette approche conduit à éliminer les propositions faites par un extracteur qui serait plus original que les autres.

Nous avons récupéré la liste des candidats-termes absents de la référence et retrouvés sur chaque corpus par exactement n extracteurs (n varie de 2 à 7 pour le corpus de brevets et de 2 à 4 pour le corpus des résumés qui n'a été traité que par quatre extracteurs). La distribution est présentée dans le tableau 1.

Filtrer par la distance au référentiel Nous faisons également l'hypothèse que les candidats-termes proposés ont plus de chance d'être valides s'ils sont proches des termes du référentiel source. Nous avons testé cette hypothèse en utilisant la distance terminologique présentée dans (Zargayouna et Nazarenko, 2010) et implémentée dans l'outil Termometer [9]. C'est une distance indépendante de la langue, qui se mesure sans faire appel à une quelconque ressource

7. ValiTerms ne nécessite pas d'installation sur le poste client mais permet d'enregistrer les validations intermédiaires en local (http ://sourceforge.net/projects/valiterms).

8. Le choix « douteux » est un ajout récent qui n'a pas été utilisé dans l'expérience relatée dans cet article.

9. http ://sourceforge.net/projects/termometerxd

Retrouvés par exactement	CB	CR
7 systèmes	89	
6 systèmes	363	
5 systèmes	1 700	
4 systèmes	12 164	3 439
3 systèmes	42 296	25 445
2 systèmes	137 114	74 576

TABLE 1 – Distribution des candidats termes absents de la référence

linguistique et qui prend en compte la compositionnalité des termes en combinant une distance sur les chaînes de caractères et une distance sur les mots.

3.2 Échantillon et résultats

Pour valider ces hypothèses, nous avons constitué un jeu de test de 3 000 candidats-termes à valider (1 500 par corpus), en équilibrant les termes retrouvés par n systèmes exactement (avec $n \geq 2$), en assurant la représentation des différents extracteurs et prenant des termes à la fois proches et éloignés de la référence selon la mesure de distance utilisée.

Nous avons donné ces 3 000 candidats-termes à valider à deux experts de l'INIST [10]. Les résultats globaux sont présentés dans le tableau 2. La première partie de ce tableau présente la proportion de termes jugés pertinents par les experts parmi les termes qu'ils ont eu à valider. La deuxième partie étudie les commentaires. Il a été demandé aux experts d'indiquer en commentaire la forme correcte des termes rejetés comme non pertinents. Les termes rejetés peuvent être mal formés ou mal orthographiés. D'autres sont des termes longs qui coordonnent plusieurs notions, dans ce cas l'expert devait indiquer le ou les sous-termes à retenir. Enfin, certains n'appartiennent pas au domaine. On constate qu'un terme, même s'il est jugé « non-pertinent », peut être intéressant à proposer à la validation parce qu'il suggère d'autres termes aux spécialistes du domaine. La dernière partie du tableau présente les termes à ajouter dans la terminologie destinée à l'annotation sémantique [11], cela correspond à l'union des termes pertinents et des termes des commentaires ne figurant pas dans le référentiel de départ.

3.3 Analyse

L'analyse de ces résultats permet de confirmer nos deux hypothèse initiales.

Le vote des systèmes et le jugement des experts sont corrélés. L'histogramme de gauche sur la figure 2 présente la proportion de termes pertinents parmi ceux qui sont retrouvés par exactement n systèmes pour les corpus de brevets (en bleu) et de résumés (en rouge). On observe que cette proportion décroît avec le nombre de systèmes [12].

10. Nous tenons à remercier Anne Busin et Marie-Pierre Verdier, spécialistes du domaine de la pharmacologie et chargées de l'indexation des articles scientifiques, pour leur travail de validation des terminologies.

11. Nous n'avons pas encore le bilan des termes à ajouter au référentiel d'indexation qui a vocation à être plus réduit que la terminologie.

12. L'histogramme bleu comporte une valeur aberrante pour 4 systèmes, qui est probablement due à une irrégularité dans la constitution du jeu de test.

	CB	CR
Termes à valider	1 500	1 500
Termes pertinents	263 (17,5%)	312 (20,8%)
Termes non pertinents	1 237 (82,5%)	1 188 (79,2%)
Termes avec un commentaire	664 (53,7%)	829 (69,8%)
Termes proposés dans les commentaires	706	941
–> qui existent déjà dans la référence	422	547
–> qui n'existent pas dans la référence	284	394
Termes à ajouter au référentiel	547 (36,5%)	706 (47,1%)

TABLE 2 – Résultats de la campagne d'enrichissement

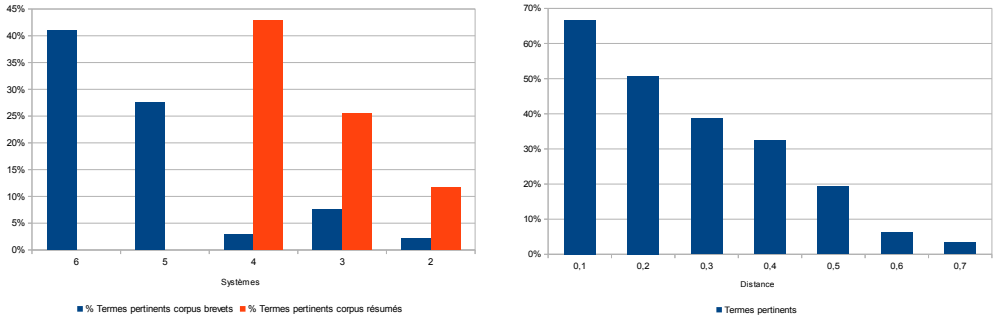

FIGURE 2 – Corrélation du nombre de systèmes (à gauche) ou de la distance (à droite) avec les jugements de pertinence

La distance terminologique et le jugement des experts sont également corrélés. Le graphique de droite sur la figure 2 montre que la proportion de termes pertinents (en ordonnée) décroît également quand la distance des termes avec ceux de la référence (abscisse) augmente. Plus les termes sont proches du référentiel (au sens de la distance terminologique), plus ils tendent à être jugés pertinents par les experts. Sur cet échantillon, si on n'avait retenu que les candidats-termes dont la distance est inférieure à 0,4, nous aurions retrouvé près de 75% de l'ensemble des termes pertinents et les experts auraient retenu près de la moitié des termes à valider comme pertinents.

Les observations faites dans le cadre de cette expérience montrent que l'on peut filtrer efficacement les candidats-termes qui sont donnés à valider à des terminologues en exploitant les sorties de différents extracteurs et/ou en s'appuyant sur une terminologie source. Le but est de donner des listes suffisamment filtrées pour que le travail de validation ne soit pas trop fastidieux et que les termes pertinents ne soient pas noyés sous le bruit. Nous considérons que juger 1 terme pertinent sur 3 constitue une tâche de validation raisonnable, d'autant que les termes rejetés en suggèrent souvent d'autres plus pertinents.

	Phase 1	Phase 2		Phase 1	Phase 2
Percent Agreement	80%	88,4%	N Accords	200	221
Pi de Scott	0,531	0,751	N Désaccords	50	29
Kappa de Cohen	0,532	0,752			

TABLE 3 – Évolution des accords inter-annotateurs

4 Contrôler la qualité de la validation

Une fois que la liste de candidats-termes à valider par les experts est constituée (à l'aide des stratégies de filtrage présentées dans la section précédente) peut débuter la phase de validation manuelle. La principale difficulté que soulève cette phase tient à la subjectivité des jugements de pertinence des experts du domaine qui est elle-même liée à leur compréhension de l'application visée et du type de terminologie que l'on cherche à construire. Par exemple un terme long comme *aerosol of stable radioactive nanoparticle* semble bien formé mais est-il pertinent pour enrichir le référentiel d'indexation, et si ce n'est pas le cas quel sous-terme privilégier ? *aerosol*, *radioactive nanoparticle* ou *stable nanoparticle* ?

Pour contrôler la subjectivité des jugements, nous proposons, en nous inspirant de la méthodologie proposée par (Fort, 2012) pour l'annotation de corpus, de mettre en place une phase de pré-campagne de validation et de calculer les accords inter-juges tout au long du processus de validation. La phase de pré-campagne permet de mettre à jour un guide de validation qui fixe les consignes de validation et l'esprit dans lequel cette validation doit être faite, jusqu'à ce que les accords deviennent satisfaisants. Une fois le niveau de qualité requis atteint, la validation à grande échelle peut se faire. Pour les campagnes de grande envergure, il est probablement souhaitable de re-mesurer également à intervalle régulier les accords intra et inter-juges pour s'assurer que le processus de validation ne dévie pas.

Nous avons proposé aux deux experts de l'INIST de valider en double aveugle 250 candidats-termes choisis aléatoirement dans notre échantillon de 3 000. Nous avons ensuite calculé les accords entre leurs jugements (première colonne du tableau 3). Comme ces valeurs étaient basses, nous avons analysé en détail les cas de désaccords dans les jugements et les commentaires. Les problèmes rencontrés étaient majoritairement dus à des questions de découpage des termes longs (par exemple *corosolic acid content of banaba extract* doit être découpé en *corosolic acid*, *banaba* et *extract*), mais aussi de généricité des termes (*review paper* est incorrect car trop générique tandis que *retrospective study* est correct car important en épidémiologie) et de termes hors du domaine du référentiel (*hydroxyglitazone*). Certains cas étaient vraiment problématiques comme *streptozotocin* qui est non pertinent (composé chimique servant à induire une pathologie expérimentale), tandis que *streptozotocin induced diabetes* est pertinent (pathologie expérimentale induite par le composé chimique). Cette analyse a permis de spécifier clairement les consignes dans le guide de validation, en dissociant notamment les objectifs d'enrichissement du référentiel d'indexation et de création d'une terminologie pour l'annotation de corpus. Cette clarification a permis d'améliorer les accords sur un nouveau jeu de 250 termes validés en double aveugle (deuxième colonne du tableau 3). A partir de là, les experts ont pu valider des 2 500 candidats termes restants. Cette expérience montre qu'en procédant avec méthode, on peut contrôler la subjectivité des jugements de validation et ainsi obtenir une terminologie de bonne qualité à partir des extracteurs de termes.

5 Conclusion et perspectives

Cet article propose une méthodologie permettant d'exploiter les sorties d'extracteurs de termes pour construire ou enrichir des terminologies à un coût et avec une qualité raisonnables. On ne peut pas se contenter de donner des listes de candidats-termes à valider aux terminologues. Cela s'apparente à chercher un terme pertinent un peu à l'aveuglette dans un amas de termes bruités : le travail de validation ne peut être de bonne qualité, l'attention se relâche, les critères deviennent flous, les objectifs sont perdus de vue.

Nous avons montré qu'on peut cependant adopter des stratégies simples pour filtrer *a priori* le gros du bruit dans les listes de candidats-termes en faisant voter plusieurs extracteurs de termes et/ou en mesurant la distance des termes proposés à ceux d'une terminologie de référence prise comme point de départ. Il reste à voir comment ces deux critères peuvent être combinés pour exploiter au mieux l'expertise humaine lors de la validation.

Nous avons montré par ailleurs qu'un protocole de validation clair, avec un guide de validation et le contrôle des accords inter-juges, permet d'atteindre une bonne stabilité de validation, seule garantie de la qualité des jugements humains qui sont ainsi posés.

Références

AUBIN, S. et HAMON, T. (2006). Improving term extraction with terminological resources. *In* SALAKOSKI, T., GINTER, F., PYYSALO, S. et PAHIKKALA, T., éditeurs : *Advances in Natural Language Processing 5th International Conference on NLP, FinTAL 2006*, pages 380–387, Turku, Finland. Springer.

CHOI, F. Y. Y. (1999). A flexible distributed architecture for nlp system development and use. *In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 615–618, Stroudsburg, PA, USA. Association for Computational Linguistics.

DAILLE, B. (2003). Conceptual structuring through term variations. *In* BOND, F., KORHONEN, A., MACCARTHY, D. et VILLACICENCIO, A., éditeurs : *Proceedings of ACL 2003 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, pages 9–16.

DROUIN, P. (2006). Termhood experiments : quantifying the relevance of candidate terms. *Modern Approaches to Terminological Theories and Applications*, 36:375–391.

FORT, K. (2012). *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus*. Thèse, Université Paris-Nord – Paris XIII.

JACQUEMIN, C. et BOURIGAULT, D. (2003). Term extraction and automatic indexing. *In* MITKOV, R., éditeur : *Handbook of Computational Linguistics*, chapitre 19, pages 599–615. Oxford University press, Oxford, GB.

MONDARY, T., NAZARENKO, A., ZARGAYOUNA, H. et BARREAUX, S. (2012). The Quaero Evaluation Campaign on Term Extraction. *In The eighth international conference on Language Resources and Evaluation (LREC)*, pages 663–669, Istanbul, Turkey.

ZARGAYOUNA, H. et NAZARENKO, A. (2010). Evaluation of Textual Knowledge Acquisition Tools : a Challenging Task. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)*, pages 435–440, Valletta, Malte.