

L'enrichissement terminologique d'usage du projet iMediate : une collaboration tripartite terminologie/TALN/sciences de la santé

Nathalie Lemaire¹, Thomas François²,
Jean-Claude Debongnie^{2,4}, Damien De Meyere², Benjamin Fauquert³, Thierry Klein³,
Cédric Fairon², Marc Van Campenhoudt¹

¹ TERMISTI, Département de traduction et interprétation, Université libre de Bruxelles

² CENTAL, IL&C, UCLouvain

³ École de Santé publique, Université libre de Bruxelles

⁴ Clinique Saint-Pierre

Nous nous proposons de partager une expérience de collaboration tripartite terminologie/traitement automatique des langues naturelles (TALN)/sciences de la santé développée dans le cadre du projet de recherche iMediate (*Interoperability of Medical Data through Information Extraction and Term Encoding*). Cette recherche vise à mettre au point des technologies pour réaliser automatiquement une représentation structurée des données « patients » à partir de textes hospitaliers non structurés (ex. : rapports de sortie d'hospitalisation) et de données cliniques structurées (ex. : analyses sanguines). Les prototypes développés visent en fait à améliorer l'interopérabilité entre les systèmes d'information susceptibles d'intervenir dans la communication entre les différents acteurs intra- et extra-hospitaliers autour des trajets de soins des patients : médecine générale, financement des séjours hospitaliers, constitution de cohortes interhospitalières de patients pour la recherche clinique, etc.

Le partenaire terminologique du projet intervient en amont de la structuration de documents médicaux jusque-là peu ou pas structurés. Son rôle est, par repérage des variations observables dans l'usage, d'enrichir les ressources terminologiques qui seront exploitées par les trois applications – annotateur, catégoriseur et moteur de recherche – développées en aval du projet. Les nombreuses nomenclatures développées par le monde médical (SNOMED CT, ICD, ICPC...) ne suffisent en effet pas à rendre compte de la variation rencontrée dans les textes médicaux « libres », qui représentent pourtant encore environ 70 % des dossiers médicaux. C'est que ces ressources de référence ne sont pas toutes intégralement traduites en français et qu'en tout état de cause, même traduites, elles privilégient la logique et la cohérence de la classification conceptuelle au détriment d'une description complète de la variation de la dénomination linguistique.

L'essentiel du travail d'enrichissement terminologique d'usage consiste donc à relever, dans un corpus de textes médicaux libres, les variations d'usage et à les « raccrocher », par regroupement synonymique, aux entrées des nomenclatures officielles. Ce travail est traditionnellement effectué par le terminologue de manière semi-automatique avec l'assistance d'outils TALN prêts à l'emploi (essentiellement extracteur terminologique et concordancier) et se clôture par la validation d'experts du domaine. Mais la taille du corpus iMediate constitué par les dossiers médicaux anonymisés des trois sponsors hospitaliers du projet confine au traitement de données massives (*big data*). Si bien qu'une adaptation de la méthodologie classique d'exploitation de corpus a été nécessaire pour que les tâches manuelles de discrimination de la pertinence des candidats termes et de validation du raccrochage conceptuel par les experts médecins soient humainement réalisables dans les limites de temps et d'effectifs disponibles. Plutôt qu'un traitement séquentiel des unités terminologiques, les équipes terminologie/TALN/sciences de la santé ont ainsi élaboré une méthodologie par cycles d'amélioration du processus, où les expertises sont, après chaque traitement de lot, appelées à rétroagir les unes sur les autres

pour une optimisation de traitement des données suivantes. Dans cette communication, nous décrivons cette méthodologie d'enrichissement terminologique d'usage, fortement basée sur la collaboration interdisciplinaire, après l'avoir succinctement replacée dans son contexte : celui des acteurs et des objectifs du projet iMediate.