Traitement Automatique de la Langue pour l'analyse rétrospective de la prise en charge des patients sur le long terme

Problématique

Avec l'avènement du dossier électronique patient, les hôpitaux génèrent des quantités de données et d'informations très importantes : typiquement, plusieurs centaines de milliers de documents par an pour un hôpital parisien. Ce vivier de données permet d'envisager des études poussées sur des cohortes de patients, afin de caractériser leur profil phénotypique, leur prise en charge et de découvrir de nouveaux biomarqueurs.

Ce projet a pour objectif de contribuer à l'analyse rétrospective de la prise en charge des patients sur le long terme. Les enjeux cliniques sont d'une part de caractériser les patients présentant un même profil diagnostique et d'autre part de mieux comprendre le parcours hospitalier des patients afin d'améliorer les protocoles de prise en charge et le suivi des patients. L'analyse s'appuiera sur les données présentes dans les dossiers électroniques des patients hospitalisés, en particulier les documents rédigés en langue naturelle, qui renseignent chaque étape du parcours d'un patient dans l'hôpital, rapportent les résultats d'analyses biologiques ou d'imagerie et fournissent également des éléments sur l'historique médical préalable du patient. Nous aurons ainsi pour but de *reconstituer automatiquement la chronologie des événements* ayant trait aux problèmes médicaux des patients, et d'associer à chaque étape de ce parcours des éléments phénotypiques pertinents.

L'originalité de ce projet repose sur une analyse fine des textes contenus dans les dossiers patient pour en extraire des évènements médicaux reliés à des marqueurs temporels, combinée à la fouille de données sur des éléments structurés. Le projet est également novateur et ambitieux dans la mesure où il propose une analyse multi-document (avec le recoupement des informations extraites de différents documents) et multi-modale (avec le recoupement des informations extraites des textes et des données structurées). Ce travail permettra, par un avancement méthodologique significatif en traitement automatique de la langue biomédicale en termes d'analyse temporelle, de faciliter et de diversifier les études rétrospectives.

Ainsi, ce projet propose de développer des méthodes et outils permettant le traitement des données hétérogènes contenues dans les dossiers patients afin de réaliser des analyses rétrospectives (« post-analyse ») permettant d'isoler des évènements médicaux et des caractéristiques phénotypiques saillantes (« facteurs influants »). Les méthodes développées pourront être utilisées de manière prédictive afin de guider l'analyse des dossiers patients et d'isoler certains profils jugés particulièrement intéressants par les praticiens.

Etat de l'art

Le traitement automatique de la langue naturelle est un outil méthodologique puissant pour analyser les documents cliniques et contribuer au processus d'aide à la décision (Demner-Fushman et al. 2009). La détection automatique de concepts médicaux a été largement étudiée, avec des résultats concluants (Savova et al. 2010). La détection des événements et des relations temporelles, que ce soit dans le domaine général (UzZaman, 2013) ou biomédical (Uzuner et al. 2013; Sun et al. 2013), est un problème complexe encore non résolu. Nous proposons de l'aborder en étant guidés par un cadre applicatif concret, nous permettant de mieux cerner les types d'événements ciblés et de bénéficier des ressources de normalisation telles que l'UMLS (Bodenreider, 2004). Ces principes nous permettent d'envisager l'étude de la coréférence événementielle multi-document, très peu abordée jusqu'à présent dans le domaine du texte.

Méthodologie

L'analyse temporelle de textes d'information a pour but général de mieux localiser dans le temps les événements décrits dans ces textes, et donc d'alimenter de façon plus précise des outils d'extraction d'information. Pour cela, la première étape est de détecter correctement les expressions temporelles de ces textes. Ces expressions temporelles peuvent être des dates absolues, c'est-à-dire que l'on peut placer sans ambiguïté sur l'axe des temps (par exemple, *le 14 janvier 2008*), mais aussi des dates relatives, qui nécessitent une phase de *résolution* ou de *normalisation* (par exemple, *le 14 janvier dernier*, *dans 6 semaines*). Dans le cadre du dossier électronique patient, des expressions temporelles propres au domaine de spécialité peuvent également être rencontrées (par exemple, *à 18 semaines d'aménorrhée, à j+1 après l'opération*). D'autre part, les types d'événements sont plus circonscrits que dans le domaine général : il s'agit par exemple d'examens, de médicaments administrés, d'interventions chirurgicales.

Les techniques d'analyse temporelle des textes ont fortement progressé ces dernières années. Elles s'attachent en général au domaine journalistique et particulièrement dans un cadre mono-document. Dans notre cas, les textes sont spécialisés, et un même événement de l'historique médical du patient peut être mentionné à plusieurs reprises dans divers documents. Ce dernier point est à la fois un facteur de complication et une grande aide pour l'élaboration des chronologies.

La difficulté est en effet de détecter les coréférences événementielles multi-documents, c'est-à-dire les mentions d'un même événement dans des textes différents, formulés de manière très variable. Du point de vue de la fouille de textes, il est intéressant d'aborder ce problème, encore hors de portée dans le domaine général, en étant guidés par un cadre applicatif concret. Les documents utilisés seront un corpus de dossiers patients recevant une chimiothérapie pour un cancer du colon.

Les avantages de la redondance pour le système d'extraction d'information que nous souhaitons mettre en œuvre se situent à deux niveaux. D'une part, l'analyse de textes étant toujours imparfaite, la répétition des informations sous des formes différentes permet d'augmenter les chances de détecter les événements. D'autre part, des événements fréquemment mentionnés dans les dossiers peuvent être considérés comme importants, ou centraux, tandis que d'autres pourront être plus anecdotiques, voire n'avoir aucun lien avec la raison principale du suivi du patient.

Résultats attendus

1. Étude sur corpus des phénomènes de coréférence événementielle dans les dossiers patients.

2. Évaluation et adaptation des outils d'analyse temporelle existants

3. Normalisation des désignations d'événements dans les textes, c'est-à-dire association entre ces désignations et les instances de ces événements dans les bases de connaissance du domaine (UMLS).

4. Analyse des liens temporels entre les événements identifiés, estimation de l'importance de ces événements pour le suivi du patient, recherche d'évènements au sein ou en dehors de l'épisode "cancer"

5. Construction automatique d’une chronologie patient à partir des différents documents constituant le dossier en cancérologie : extraction des séquences temporelles complexes des chimiothérapie (protocoles, cycles, cures, jours) et évènements intercurrents.

6. Évaluation et confrontation des résultats avec le protocole standard, en collaboration avec nos partenaires hospitaliers : on s’attachera notamment à identifier les décalages entre les

séquences de chimiothérapie et au sein des séquences, les substitutions au sein des séquences

Contact

Aurélie Névéol et Xavier Tannier
{neveol;xtannier}@limsi.fr
LIMSI-CNRS
Rue John von Neumann
Université Paris-Sud
91403 ORSAY
France

Références

Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-70.

Demner-Fushman D, Chapman WW, McDonald CJ: What can natural language processing do for clinical decision support? J Biomed Inform 2009, 42:760–772.

Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES: architecture, component evaluation and applications. J Am Med Inform Assoc. 2010 Sep-Oct;17(5):507-13.

Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inform Assoc. 2013 Sep-Oct;20(5):806-13.

Uzuner O, Stubbs A, Sun W. Chronology of your health events: Approaches to extracting temporal relations from medical narratives. J Biomed Inform. 2013 Dec;46 Suppl:S1-4.

UzZaman N, Llorens H, Derczynski L, Verhagen M, Allen J, Pustejovsky J. SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations Proceedings and the Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013), ACL, 2013, 1-9