PROPOSITION DE SUJET DE THESE

NOM DU LABORATOIRE : LIMICS, LABORATOIRE D'INFORMATIQUE MEDICALE ET D'INGENIERIE DES CONNAISSANCES EN E-SANTE
DIRECTEUR DE THESE : XAVIER TANNIER
ADRESSE : 15 RUE DE L'ECOLE DE MEDECINE, 75006 PARIS

TITRE DE LA THESE : Enrichissement des données d'un entrepôt de données de santé par annotation précise de documents cliniques textuels par apprentissage semi-supervisé

CO-ENCADRANTE : CHRISTEL DANIEL
EQUIPE DE LA CO-ENCADRANTE : AP-HP WIND, LIMICS
LABORATOIRE : LIMICS

PRESENTATION DU SUJET

Mots-clés : extraction d'information dans les textes, apprentissage faiblement supervisé, réseaux de neurones, traitement automatique des langues, entrepôt de données de santé

Contexte

Présents en très grande quantité dans les entrepôts de données de santé, les documents cliniques hospitaliers (comptes rendus d'hospitalisation ou de consultation, comptes rendus opératoires, de radiologie, d'anatomie pathologie, transmissions infirmières, lettres et ordonnances de sortie, ou courriers de médecins, par exemple) constituent de riches sources d'informations pour diverses applications telles que le recrutement de patients pour la recherche clinique, la surveillance épidémiologique, le codage médical et les outils d'aide à la prise de décision [Wang2018]. La détection des concepts médicaux (maladies, signes, symptômes, traitements, médicaments, etc.) manipulés dans ces documents est donc un sujet de recherche important en traitement automatique des langues [Liu2016]. Ces documents, écrits en langage naturel, par des humains et pour des humains, sont encore très difficiles à analyser et donc à valoriser, en raison de la variation du langage en général, mais également du caractère technique des documents, dont le vocabulaire varie très fortement d'une spécialité médicale à une autre. Par ailleurs, outre l'extraction des mentions des concepts médicaux, il est crucial de pouvoir les caractériser, en particulier de déterminer s'ils sont niés, incertains, d'intensité forte ou faible, ou se rapportent à une autre personne que le patient (membre de la famille ou tiers) puisque ces aspects peuvent bien entendu changer radicalement la compréhension du texte. Enfin, pour les langues autres que l'Anglais, le manque d'outils, de données et de ressources terminologiques est également un frein important. Par exemple, en Français, il n'existe que deux corpus cliniques annotés qui couvrent un petit sous-ensemble du domaine médical [Campillos2018, Lerner2019], et les ontologies internationales telles que le système de langage médical unifié (UMLS®) ne sont pas entièrement traduites [Névéol2014]. Le coût de développement d'un tel corpus annoté est très élevé, car il a été rapporté que l'annotation de 5 documents médicaux pour 12 entités prend en moyenne 82 minutes [Campillos2018].
Dans le domaine général (non médical), les systèmes supervisés basés sur l'apprentissage automatique se sont révélés plus efficaces que les systèmes à base de règles et de terminologies [Lample2016]. Cependant, pour les raisons citées ci-dessus, la reconnaissance des concepts médicaux sans restriction de domaine par ces méthodes demanderait un effort d'annotation hors de portée. Des travaux ont commencé à aborder la question des systèmes non supervisés [Zhang2013, Alicante2016] ou semi-supervisés [DelVigna2016, Gupta2018, Fries2017, Liu2019] pour l'Anglais.

L'objectif de cette thèse est donc d'explorer plusieurs approches de réduction de la supervision pour une annotation multilingue et généraliste (s'étendant à l'ensemble des domaines médicaux et à l'ensemble des types de documents) des dossiers cliniques. Ceci permettra d'envisager l'application d'outils d'extraction à l'ensemble des comptes-rendus d'un entrepôt de données de santé (par exemple, 50 millions actuellement à l'AP-HP), et ainsi que collecter et de structurer de très importantes quantités d'information qui restait jusqu'ici inexploitables. Par ailleurs, les données étant très difficilement partageables dans ce domaine, toute approche nécessitant de grandes quantités de données annotées n'est pas partageable avec la communauté pour une application dans d'autres entrepôts. Notre travail contribuera à réduire ce problème.

Objectifs et méthodes

Nous traiterons quatre questions distinctes liées à l'extraction d'information dans les textes cliniques :

1. L'utilisation des terminologies du domaine médical comme moyen de supervision distante pour l'extraction de concepts médicaux.
2. L'utilisation d'outils de traduction automatique des terminologies et des textes pour l'adaptation à la langue, avec des approches par transfert.
3. L'utilisation de l'apprentissage actif pour réduire l'annotation manuelle des caractéristiques des concepts (négation, contexte, incertitude, intensité)
4. Qualité des données extraites et intégration dans l'entrepôt de données

Les deux premières tâches sont deux approches d'un même problème : il s'agit de proposer un outil pouvant s'appliquer avec des performances équivalentes sur tous types de documents cliniques (résumé, lettre, ordonnance) et pour toutes les spécialités médicales. S'il ne semble pas envisageable d'obtenir des données annotées suffisamment complètes pour entraîner un modèle supervisé efficace, des premières expérimentations suggèrent en revanche que l'analyse de grandes quantités de données non annotées, avec l'aide de terminologies, peut pallier ce manque avec de bonnes performances [Liu2019, ainsi que des travaux préliminaires réalisés au LIMICS] (point 1 ci-dessus). D'autre part, la traduction de termes et de textes, y compris sans supervision, ayant fait des progrès spectaculaire ces dernières années [Conneau2018], nous souhaitons utiliser les terminologies et les textes annotés en langue anglaise pour appliquer des méthodes d'apprentissage par transfert sur des textes en français, le but étant d'obtenir une approche totalement indépendante de la langue (à l'exception des ressources disponibles pour la traduction), avec l'Anglais comme langue pivot (point 2). Ce dernier point s'articule avec la thèse de Nicolas Paris actuellement en cours au LIMICS avec les mêmes encadrants.

Pour la troisième tâche, nous pensons que les caractéristiques des concepts comme la négation, le contexte, l'incertitude ou l'intensité des phénomènes dépendant trop de la syntaxe et des expressions spécifiques à une langue pour bénéficier des mêmes approches avec succès. En revanche, la variation sur les formes d'expression de la négation, par exemple, ne demande pas une quantité de données annotées hors de portée, et est relativement indépendante de la spécialité médicale traitée (les concepts sont différents mais les expressions utilisées pour la négation sont similaires). D'autre part, la portée de ces expressions est locale (dépasse rarement le cadre de la phrase), et l'annotation manuelle en est simplifiée puisqu'elle ne nécessite pas la lecture attentive d'un cas entier, mais plutôt une lecture phrase par phrase sans mémoriser le contexte. Pour toutes ces raisons, le recours à des approches classiques d'apprentissage actif sera ici bénéfique et permettra de réduire fortement le temps d'annotation nécessaire.

Enfin, une intégration des données extraites à l'entrepôt de données de santé peut permettre d'enrichir de façon considérable les données disponibles, et donc d'ouvrir des perspectives importantes pour la recherche et le soin. Cependant, la question de la qualité de ces données produites automatiquement devra être étudiée. Aux méthodes classiques d'évaluation de ce type de systèmes (précision et rappel), nous ajouterons une étude de l'utilisabilité des données bruitées dans le cadre de l'entrepôt.

Sources de données

Les premières expérimentations sur la langue anglaise s'appuieront sur des jeux de données librement disponibles (CLEF-Share, i2b2, MIMIC). Dans le but d'appliquer nos approches sur des données en français avec un volume plus important, une demande d'accès aux données de l'EDS de l'AP-HP sera effectuée auprès du Comité Scientifique et Éthique (CSE) de cet EDS. Cette étape est connue et maîtrisée par les encadrants.

Encadrement

Un encadrement interdisciplinaire est indispensable à ce type de sujet. Xavier Tannier possède une expertise sur l'extraction d'information dans les textes, avec des approches supervisées ou non supervisées [Nguyen2015] ; il a également travaillé sur les documents cliniques avec des méthodes neuronales [Tourille2017, Tourille2018]. Christel Daniel apporte à la fois son expertise médicale et ses connaissances sur l'entrepôt de données de santé de l'AP-HP et sur la qualité des données [Daniel2018].

Références

[Alicante2016] Alicante A, Corazza A, Isgrò F, Silvestri S. Unsupervised entity and relation extraction from clinical records in Italian. Comput Biol Med. 2016;72: 263–275.

[Campillos2018] Campillos L, Deléger L, Grouin C, Hamon T, Ligozat A-L, Névéol A. A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT). Language Resources and Evaluation. 2018;52: 571–601.

[Conneau2018] Conneau, A.; Lample, G.; Ranzato, M.; Denoyer, L. & Jégou, H. Word translation without parallel Data Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018), 2018

[Daniel2018] Daniel C, Serre P, Orlova N, Bréant S, Paris N, Griffon N. Initializing a hospital-wide data quality program. The AP-HP experience. Comput Methods Programs Biomed. 2018 Nov 9. pii: S0169-2607(18)30624-2.

[DelVigna2016] Del Vigna F, Petrocchi M, Tommasi A, Zavattari C, Tesconi M. Semi-supervised Knowledge Extraction for Detection of Drugs and Their Effects. Social Informatics. Springer International Publishing; 2016. pp. 494–509.

[Fries2017] Fries J, Wu S, Ratner A, Christopher R. SwellShark: A Generative Model for Biomedical Named Entity Recognition without Labeled Data [Internet]. arXiv [cs.CL]. 2017.

[Gupta2018] Gupta S, Pawar S, Ramrakhiyani N, Palshikar GK, Varma V. Semi-Supervised Recurrent Neural Network for Adverse Drug Reaction mention extraction. BMC Bioinformatics. 2018;19: 212.

[Lample2016] Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural Architectures for Named Entity Recognition. In proceedings of the North American Chapter of the ACL (NAACL-HTL 2016). San Diego, USA, 2016.

[Lerner2019] Lerner Y, Paris N, Tannier X. Terminologies augmented recurrent neural network model for clinical named entity recognition. https://arxiv.org/pdf/1904.11473.pdf

[Liu2016] Liu F, Chen J, Jagannatha A, Yu H. Learning for Biomedical Information Extraction: Methodological Review of Recent Advances [Internet]. arXiv [cs.CL]. 2016.

[Liu2019] Liu, A.; Du, J. & Stoyanov, V. Knowledge-Augmented Language Model and its Application to Unsupervised Named-Entity Recognition. Proceedings of the North American Chapter of the ACL (NAACL-HLT 2019), 2019

[Névéol2014] Névéol A, Grosjean J, Darmoni SJ, Zweigenbaum P, Others. Language Resources for French in the Biomedical Domain. LREC. 2014. pp. 2146–2151.
[Nguyen2015] Kiem-Hieu Nguyen, Xavier Tannier, Olivier Ferret, Romaric Besançon. Generative Event Schema Induction with Entity Disambiguation. in Proceedings of the 53rd ACL meeting. Beijing, China, July 2015
[Wang2018] Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: A literature review. J Biomed Inform. 2018;77: 34–49.
[Zhang 2013] Zhang S, Elhadad N. Unsupervised biomedical named entity recognition: experiments with clinical and biological texts. J Biomed Inform. 2013;46: 1088–1098.

PREREQUIS, FORMATION : FORMATION D'INFORMATIQUE, CONNAISSANCES AVANCEES EN APPRENTISSAGE AUTOMATIQUE ET EN TRAITEMENT AUTOMATIQUE DES LANGUES

CONTACT POUR CE SUJET : XAVIER TANNIER, CHRISTEL DANIEL
EMAIL : XAVIER.TANNIER@SORBONNE-UNIVERSITE.FR, CHRISTEL.DANIEL@APHP.FR
TELEPHONE : 01 44 27 91 13