Élémentaire mon cher Watson?

Jean Charlet, Xavier Tannier

Élémentaire mon cher Watson?

Jean Charlet[1,2], Xavier Tannier[1]

[1] Sorbonne Université, INSERM, Université Sorbonne Paris Nord, UMR_S 1142, LIMICS, Paris
[2] Assistance Publique-Hôpitaux de Paris, DRCI, Paris, France
prenom.nom@sorbonne-universite.fr

Résumé : Le système Watson de IBM fait le buzz depuis quelques années. Ce buzz n'est pas toujours à l'avantage du système, en particulier en médecine où un article de Stat News de février 2017 met en avant l'échec de Watson. Dans cet article, nous tentons d'analyser cet échec et le comparer à ce que l'on peut attendre des systèmes d'IA, en particulier par rapport aux techniques mises en œuvre dans Watson en termes de traitement automatique du langage en médecine.

IBM's Watson system has been buzzing for a few years. This buzz is not always to the benefit of the system, especially in medicine, where a Stat News article from February 2017 highlights Watson's failure. In this article, we try to analyze this failure and to compare Watson to what can be expected from AI systems, in particular in terms of natural language processing in medicine.

Mots-clés : TALM, Ontologies, Apprentissage.

1 Introduction

Le système Watson de IBM fait le buzz depuis quelques années. Ce buzz n'est pas toujours à l'avantage du système, en particulier en médecine où un article de Stat News de février 2017 met en avant l'échec de Watson et rapporte comment le MD Anderson Cancer Center (MDACC) a dépensé plus de 60 millions de dollars avant de cesser tout travail autour du système [1]. Dans cet article, nous voudrions essayer d'analyser cet échec et le comparer à ce que l'on peut attendre des systèmes d'IA, en particulier par rapport aux techniques mises en œuvre dans Watson en termes de reconnaissance du langage écrit en médecine, que l'on appelle traitement automatique du langage médical (TALM).

Dans la section 2 nous allons décrire le principal contexte d'utilisation du TALM, à savoir les entrepôts de données de santé des hôpitaux. Dans la section 3, nous redonnons une rapide définition de l'IA. Dans les sections 4 et 5, nous développons les principales difficultés que rencontre le TALM. Dans la section 6, nous analysons autant que faire ce peut le fonctionnement de Watson. Dans la section 7, nous présentons des projets du LIMICS exemplaires de notre façon d'aborder le TALM et en notant les limites des travaux. Enfin, dans la section 8, nous discutons des résultats obtenus par Watson et de la difficulté intrinsèque de la tâche qui lui est dévolue.

2 Structurer et prédire

En quelques années, l'organisation des données dans les hôpitaux a évolué fondamentalement : Il a été acté qu'il fallait séparer les bases des données patients liées aux soins de bases liées à la recherche qui récupèrent les mêmes données pour les mettre dans des formats facilitant leur traitement : les entrepôts de données cliniques. Accessoirement, cela permet aussi d'interroger la 2e base sans empiéter sur la première qui assure la continuité des données liées au soin. En passant, l'efficacité de ces entrepôts en termes de représentation des données, de mises à jour et de temps de réponse fait qu'ils commencent aussi à être utilisés pour le soin : ils servent par exemple à croiser rapidement des données sur des groupes de patients pour analyser ou optimiser des traitements.

Ainsi, les entrepôts de données cliniques se répandent en France et dans le monde, rassemblant une grande quantité de données sur les parcours des patients à l'hôpital (actuellement

1. https://www.statnews.com/2017/09/05/watson-ibm-cancer/

50 millions de rapports à l'AP-HP). De tels volumes ouvrent de vastes perspectives d'applications nouvelles pour le soin, la recherche et le pilotage médico-économique. En particulier, la promesse d'une médecine personnalisée, guidant les médecins vers des choix thérapeutiques plus adaptés au profil des patients grâce à l'étude de larges cohortes, a motivé de nombreuses publications et de nombreux programmes de recherche.

Notamment, deux grandes classes de tâches ont émergé ces dernières années en ce qui concerne l'utilisation automatique et massive des documents hospitaliers pour la médecine personnalisée : d'une part, la structuration d'informations présentes de façon non structurée dans les dossiers patients, et d'autre part la prédiction d'événements en fonction des caractéristiques propres à un patient. Cette prédiction peut concerner la réponse à un traitement ou la survenue d'un problème (hospitalisation, rechute, décès…).

3 Qu'est-ce que l'IA

Nous commençons par un rapide point sur ce qu'est l'IA pour positionner un peu le contexte de cet article. L'intelligence artificielle est née dans les années 1950 avec l'objectif de faire produire les tâches humaines par des machines mimant l'activité du cerveau. Face aux déboires des premières heures, deux courants se sont constitués. Les tenants de l'intelligence artificielle dite forte visent à concevoir une machine capable de raisonner comme l'humain, avec le risque supposé de générer une machine supérieure à l'homme et dotée d'une conscience propre. Cette voie de recherche est toujours explorée aujourd'hui, même si de nombreux chercheurs en IA estiment qu'atteindre un tel objectif est impossible à moyen terme. D'un autre côté, les tenants de l'intelligence artificielle dite faible mettent en œuvre toutes les technologies disponibles pour concevoir des machines capables d'aider les humains dans leurs tâches. Ce champ de recherche mobilise de nombreuses disciplines, de l'informatique aux sciences cognitives en passant par les mathématiques, sans oublier les connaissances spécialisées des domaines auxquels on souhaite l'appliquer. Ces systèmes, de complexité très variable, ont en commun d'être limités dans leurs capacités ; ils doivent être adaptés pour accomplir d'autres tâches que celles pour lesquelles ils ont été conçus.

3.1 Certains systèmes d'IA utilisent la logique…

L'approche la plus ancienne historiquement s'appuie sur l'idée que nous raisonnons en appliquant des règles logiques (déduction, classification, hiérarchisation, …). Les systèmes conçus sur ce principe appliquent différentes méthodes, qu'elles soient fondées sur l'élaboration de modèles d'interaction entre agents (systèmes multi-agents), de modèles syntaxiques et linguistiques (traitement automatique des langues) ou d'élaboration d'ontologies (représentation des connaissances). Ces modèles peuvent être utilisés ensuite par des systèmes de raisonnement logique pour produire des faits nouveaux. L'approche, dite symbolique, a permis le développement, dans les années 1980, d'outils capables de reproduire les mécanismes cognitifs d'un expert, et baptisés pour cette raison systèmes experts. Les difficultés de modélisation des connaissances ont amené un certain échec de ces systèmes. Actuellement, des systèmes dits « d'aide à la décision » sont développés : Ils bénéficient de meilleurs modèles de raisonnement ainsi que de meilleures techniques de description des connaissances médicales, des patients et des actes médicaux. De plus, ils ne cherchent plus à remplacer le médecin mais à l'épauler dans un raisonnement fondé sur les connaissances médicales de sa spécialité. Ces systèmes permettent aussi d'effectuer des tâches de pilotage de systèmes multi-agent ou de traitement automatique des langues, etc. Nous sommes dans l'IA symbolique.

3.2 ... D'autres exploitent l'expérience passée...

Contrairement à l'approche symbolique, l'approche dite numérique raisonne sur les données. Le système cherche des régularités dans les données disponibles pour extraire des connaissances, sans modèle préétabli. Cette méthode née dans les années 1980 s'est popularisée depuis le début des années 2000 grâce à l'augmentation de puissance des ordinateurs

et à l'accumulation des gigantesques quantités de données qu'il est convenu d'appeler méga données ou big data.

Une majorité des systèmes actuels procèdent par apprentissage automatique, une méthode fondée sur la représentation informatique et statistique de situations existantes et connues, dans le but d'apprendre à généraliser à des données nouvelles. La force de cette approche est que l'algorithme apprend la tâche qui lui a été assignée par essais et erreurs, avant de se débrouiller tout seul. De tels systèmes s'attaquent aux mêmes problématiques que l'IA symbolique avec des résultats parfois plus probants : des applications existent en aide à la décision, en traitement automatique des langues, etc. L'apprentissage profond a notamment obtenu des résultats significatifs en analyse d'images, par exemple pour repérer, sur les photos de peau, de possibles mélanomes, ou bien pour détecter des rétinopathies diabétiques sur des images de rétines. Leur mise au point nécessite de grands échantillons d'apprentissage : 50 000 images pour le mélanome, 128 000 pour la rétine sont nécessaires pour entraîner l'algorithme à identifier les signes de pathologies. Pour chacune de ces images, on lui indique à quel ensemble elle appartient. À la fin de l'apprentissage, l'algorithme arrive à reconnaître avec une excellente performance de nouvelles images présentant une anomalie.

3.3 ... Mais les 2 visent les mêmes buts

Pour les sujets qui nous intéressent, les 2 approches proposent des solutions, en particulier pour la structuration des textes des dossiers patients : on parle d'annotation sémantique en IA symbolique et on utilise pour cela des Systèmes d'Organisation des Connaissances (SOC) divers (ontologies, classifications, etc.) et d'apprentissage de modèles patient pour l'IA numérique.

4 Le problème de la reproductibilité

Si la littérature sur ces sujets est vaste, une mise en production efficace et générale de systèmes automatiques dans les hôpitaux ou au service d'une système de santé tarde à se mettre en place, pour des raisons diverses liées à la problématique générale de la reproductibilité des approches employées. Les problèmes habituels de variabilité des données sont en effet, dans le domaine clinique, accentués par de nombreux paramètres :

- nature technique des documents et nombre élevé de spécialités médicales, conduisant à un vocabulaire pléthorique,
- faible niveau de normalisation des systèmes d'information et des terminologies utilisées dans les hôpitaux,
- hétérogénéité des natures de données : texte, image, données numériques (résultats d'analyse), séries temporelles (EEG, ECG...), données omiques,
- hétérogénéité des sources d'information : appareils de mesures, lettres, ordonnances, rapports, etc.
- utilisation des langues locales dans le cadre du soin mais de l'anglais en recherche.

Ainsi, des systèmes conçus ou des modèles entraînés sur certains types de données s'avèrent souvent inefficaces lors de leur application à un problème similaire sur des données légèrement différentes.

Enfin, le caractère hautement confidentiel des données manipulées empêche le partage entre les différents acteurs, freinant d'une part les initiatives structurantes autour d'une communauté comme d'autres domaines ont pu le vivre, et limitant fortement la reproductibilité et la comparaison des approches, en l'absence de benchmark commun. Ce problème est particulièrement bloquant dans le domaine du traitement automatique des langues, les documents textuels contenant un grand nombre d'informations sensibles et identifiantes difficiles à anonymiser.

5 Une difficile adaptation au domaine

Malgré le souhait de structurer les dossiers des patients à la source, plus de 80% des données hospitalières sont collectées sous forme de textes, principalement dans des comptes rendus cliniques. Ces documents, écrits en langage naturel, par des humains et pour des humains, sont encore très difficiles à analyser et donc à valoriser. Cela tient à la variation de la langue en général, mais aussi à la nature technique des documents, dont le vocabulaire varie fortement d'une spécialité médicale à l'autre. Il est très difficile d'extraire de ces textes une valeur informative exploitable, telle que des antécédents personnels et familiaux, un mode de vie, des symptômes, des signes, des diagnostics, des actes, des résultats d'analyses biologiques ou d'imagerie, des traitements médicamenteux ou non. Une fois extraites, un autre défi consiste à les mélanger avec les données structurées disponibles, afin d'obtenir une représentation complète du patient. Enfin, un dernier sujet consiste à interroger ces concepts et représentations afin de rechercher des patients présentant des caractéristiques données (phénotypage) ou de récupérer des cas médicaux similaires.

Toutes les tâches liées au traitement automatique des textes cliniques sont touchées par la difficulté d'adapter des systèmes à des corpus ou à des domaines différents, et a fortiori à des langues différentes (Névéol *et al.*, 2018). Même les tâches les plus simples, qui semblent a priori indépendantes du domaine, et qui sont parfois à tort considérées comme résolues, sont concernées :

— segmentation en mots et en phrases (Tapi Nzali *et al.*, 2015) ;
— gestion de la négation dans les textes (Wu *et al.*, 2014) ;
— détection des expressions temporelles (Strötgen & Gertz, 2013; Nzali *et al.*, 2015).

Ce constat se trouve aggravé pour toutes les tâches plus complexes comme la reconnaissance de concepts médicaux ou de relations (Tourille *et al.*, 2017) dans le but de la détermination des caractéristiques cliniques ou biologiques des patients, tâche ultime de la structuration des dossiers patients. Si des travaux ont montré que l'aide à la constitution de cohortes peut permettre de faciliter et d'accélérer le travail des chercheurs (Gottesman *et al.*, 2013; Wei & Denny, 2015), voire même de conduire à de nouvelles découvertes cliniques (Denny *et al.*, 2013; Carroll *et al.*, 2015; Ritchie *et al.*, 2014; Lin *et al.*, 2015), ces approches demandent un investissement de départ considérable et sont difficiles à généraliser car les annotations manuelles sont spécifiques à un cas particulier ; c'est la raison pour laquelle elles n'ont été appliquées que sur un nombre de phénotypes relativement limité, y compris dans des efforts particuliers de généralisation (Halpern *et al.*, 2016; Agarwal *et al.*, 2016; Beaulieu-Jones & Greene, 2016).

Au-delà de la question des variations linguistiques, se pose la question des disparités structurelles et conceptuelles introduites par l'utilisation de modèles de données différents dans le cadre de l'IA numérique. Si des standards de modèles de données communs émergent depuis quelques années, la migration vers ces modèles est très lente, et des études montrent que certains résultats ne sont pas reproductibles entre un modèle et un autre (Xu *et al.*, 2015) ou d'un centre hospitalier à un autre (Madigan *et al.*, 2013). Dans le cadre de l'IA symbolique, le problème est similaire : de très nombreux SOC existent et il s'en développe aussi rapidement que des applications. L'existence d'un modèle de référence – terminologie de référence (Rosenbloom *et al.*, 2006) – qui couvrirait toutes les spécialités médicales et de santé n'a pas été prouvée même si de larges initiatives existent (UMLS, SNOMED, CIM-11, NCIt). A l'inverse, des travaux proposent de faire de l'annotation sémantique en alignant des SOC au sein d'un serveur de terminologie mettant à disposition autant de ressources que nécessaire [2].

6 L'exemple de Watson

6.1 Introduction à Watson et sa communication

Il est difficile de comprendre le fonctionnement exact de Watson, d'abord parce que sous ce vocable, IBM nomme tous ses outils d'IA et qu'ensuite, société privée exige, elle n'expli-

2. https://www.hetop.eu/hetop/ avec « sélection de terminologies » (en haut à gauche de la fenêtre).

cite pas le fonctionnement des différents modules qui composent Watson. Mais grâce à cet article (Cf. *infra*, 1re URL), et à une analyse qui en est refaite dans Internetactu [3], on peut essayer d'aller plus loin. De plus, on trouve dans PubMed quelques articles écrits par des utilisateurs de Watson en médecine qui permettent de sortir des discours marketing (Simon *et al.*, 2019; Lee *et al.*, 2018). Enfin, dans une dernière page consultée le 04/03/2020 [4], on note une intéressante discussion tenue par des médecins sur les espoirs et limites de Watson [5]. Par la suite, pour plus de commodité, nous utiliserons le nom de Watson sans spécifier IBM ou système et comme si on relatait les performances d'une personne.

6.2 Le fonctionnement de Watson dans les hôpitaux

À la lecture de Stat News (additionné des réflexions de Internetactu), on comprend que Watson a été installé dans le MDACC pour tenter de colliger et d'analyser l'ensemble des données patients de l'hôpital. Quant à préciser ce qui a été exactement fait, nous nous appuyons sur l'article de The Oncologist même si les auteurs sont en lien d'intérêt avec IBM (Simon *et al.*, 2019). Par ailleurs, Watson a été utilisé en Corée, à une moindre échelle et a donné lieu à une évaluation sur la qualité des recommandations faites, avec les mêmes précautions de notre côté que pour l'autre article (Lee *et al.*, 2018). A la lecture de ces articles et des sites précédemment cités, on peut expliciter quelques tâches qui ont été effectuées par Watson :

Récapitulatifs des antécédents du patient. Une tâche importante pour décider de plans de soin difficile puisqu'elle sous-entend que l'on retrouve l'historique du patient, temporalisé. Les résultats semblent donner une F-mesure de 0,651.

Découverte du bon traitement pour le patient. L'article de (Simon *et al.*, 2019) semble trouver de bons traitements avec une concordance dans une fourchette de 0,89 à 0,99. Mais les sites montrent des résultats beaucoup plus décevants (0,33) dès qu'il faut adapter les recherches de Watson à d'autres pays, le système proposant des protocoles non réglementaires dans le pays impliqué. Dans des populations spécifiques (âgées) analysées dans l'article de (Simon *et al.*, 2019), la concordance baisse même à 0,2. Dans certains cas, le protocole proposé pourrait même être dangereux.

Fourniture des données de la littérature scientifique. Stat News signale que, en ce domaine, Watson fournit bien souvent les meilleures données de la littérature scientifique sur les traitements. Il permettrait également de mieux discuter des options possibles avec les patients et entre médecins. On peut noter que c'est une tâche plus facile car la recherche bibliographique n'est pas une aide à la décision, c'est proposer quelques liens (très) bien choisis en fonction de l'analyse en TALM du résumé des articles.

Enfin, les médecins sus-nommés dans le site « Innovation e-santé », mettent en avant, en plus des problèmes rapportés ci-dessus, le fait que la meilleure étude de médecine fondée sur les preuves, effectuée dans un endroit peut être contredite par une autre étude ayant en théorie les mêmes tenants et aboutissant. La différence est liée à des petites différentes, souvent pratiques, sur la façon de faire des procédures de soin/chirurgicales, sur l'appréhension des données, qui rend les corpus de textes comme de données hétérogènes et biaise ensuite les méthodes de travail. C'est la sempiternelle question du contexte d'explicitation des connaissances. Contexte que les systèmes d'IA ne savent pas prendre correctement en compte, Watson comme les autres.

On peut retirer de ces quelques informations que les tâches attribuées à Watson dans son ensemble sont difficiles et leur réussite au niveau d'exigence de la médecine et de façon

3. http://www.internetactu.net/a-lire-ailleurs/watson-une-revolution-pour-lutter-contre-le-cancer-nous-en-sommes-loin/

4. https://innovationesante.fr/linformatique-cognitive-de-watson-dibm-au-service-de-lhomme-watson-health-espoirs-et-limites/

5. De nouveaux « billets » sur Stat News sont très critiques par rapport aux recommandations du système (https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/) et par rapport à la mauvaise localisation des guides de bonne pratique par rapport au pays où il est déployé (https://www.statnews.com/2018/07/31/ibm-watson-modifying-cancer-treatment-software/). Mais comme ils sont en accès payants, nous ne les discutons pas ici.

concomitante est une gageure : en particulier, une décision de choix de protocole qui se fonde sur un historique mal analysé risque d'être inadéquate ; si on rajoute que la médecine admet difficilement l'erreur, on a là quelques éléments qui expliquent l'échec de Watson.

7 Au LIMICS

Dans les nombreux projets développés au LIMICS, certains ont pour objectif d'extraire des informations pertinentes des textes médicaux. Dans ce contexte, toutes les approches sont utilisées, dans le champ de l'IA symbolique, l'annotation sémantique des textes grâce à des ontologies ou des terminologies dans le champ de l'IA numérique et, enfin des approches mixtes.

Nous allons décrire des projets, exemples de chacune des approches en analysant les réussites avérées et potentielles et nous analyserons les limites en discussion.

Le projet Paron vise à analyser le parcours de soin des patients atteints de Sclérose Latérale Amyotrophique (SLA) et à expliciter les déterminants. La pathologie provoque de nombreuses incapacités et situations de handicaps et ces patients nécessitent un accompagnement pluridisciplinaire. Cette prise en charge complexe peut amener à des situations de rupture de parcours par l'absence, l'arrêt ou des difficultés de prise en charge. Cependant les causes de ces ruptures ne sont pas connues. Le réseau de coordination ville hôpital SLA Île-de-France dispose d'une base textuelle de coordination, sur laquelle les besoins et les demandes des patients sont décrits tout au long de leurs parcours. Dans cette thèse, nous proposions d'analyser cette base pour en extraire de la connaissance et décrire les parcours patients. Le domaine de la coordination des soins dans un réseau de pris en charge de maladie dégénérative étant extrêmement spécifique, il n'y a pas de Système d'Organisation des Connaissances en général ou encore mieux, d'ontologie disponible. Il a donc fallu construire une ontologie modulaire, OntoParon, recouvrant trois sous-domaines : (1) la médecine liée à la SLA, (2) la coordination des soins et (3) les concepts sociaux-environnementaux, très importants dans le domaine du handicap. En utilisant les propriétés de classifications des ontologies, des concepts définis, rendant compte de thématiques importantes comme l'épuisement de l'aidant ou bien encore la présence de problèmes sociaux, ont été créés pour détecter les difficultés rencontrées autour des patients.

Un outil d'annotation sémantique, OnBaSAM, utilise cette ontologie pour retrouver des éléments d'information importants dans les textes. La qualité des annotations est évidemment un critère majeur de la validité des analyses proposées. Leur validité, rapportée à des gold standards de professionnels donne une F-mesure de 0,96.

Ainsi, l'annotation de 931 dossiers patients permet de faire des analyses statistiques sur les données ainsi générées et montre que tous les patients ne présentent pas les mêmes besoins ni les mêmes demandes : certaines thématiques s'expriment différemment en fonction de l'âge, de la forme de pathologie ou du mode de vie des personnes. Le nombre de dossiers pris en compte (Cardoso, 2019).

Les premières étapes du projet terminé, on peut noter qu'on a des résultats statistiques intéressants à partir de résultats de TAL de (très) bonne qualité. Le crédit revient en partie à la qualité de l'ontologie. Il faut évidemment préciser, comme noté ailleurs dans le document, que le système ne marche que pour le domaine précis de l'ontologie et pour des documents de la forme de ceux pour lesquels OnBaSAM a été paramétré. On espère une certaine généricité de l'approche en modifiant l'ontologie mais cela reste à démontrer. Par ailleurs, on notera que c'est une étude épidémiologique, une intervention d'analyse « froide » au contraire de systèmes d'IA qui font – ou devront/devraient faire – de l'aide à la décision chaude, en routine.

Une autre approche d'annotation sémantique est d'utiliser plusieurs ressources sémantiques pour annoter comme l'ECMT développé par Darmoni *et al.* (2018). La difficulté réside alors dans l'investissement nécessaire pour développer le serveur y compris l'alignement des ressources entre elles pour permettre l'annotation sans développer une ressource spécifique comme dans l'exemple précédent. Inversement, l'approche est probablement plus générique qu'avec une seule ontologie. Les développements autour de l'ECMT se poursuivent avec des

techniques de *Word embedding* associées à l'annotateur (Dynomant *et al.*, 2019).

Concernant les applications de techniques d'apprentissage aux textes médicaux, la situation évoquée en introduction guide un objectif général de réduction de la supervision humaine nécessaire pour transférer les connaissances expertes vers le système, en général par le biais de données annotées, mais aussi grâce à une représentation pertinente de ces connaissances. Il s'agit donc de diminuer les efforts à mettre en œuvre pour permettre une réponse à une question médicale lorsque cette réponse nécessite l'analyse de documents textuels.

Deux approches complémentaires sont alors à l'étude. D'une part, l'annotation générique des mentions de concepts médicaux dans les textes, pour mieux caractériser de façon générale et sans but prédéfini le parcours d'un patient. Ici, il s'agit de proposer un outil de détection et de caractérisation des concepts (qui peuvent être niés ou de factualité incertaine), pouvant s'appliquer avec des performances équivalentes sur tous types de documents cliniques (résumé, lettre, ordonnance) et pour toutes les spécialités médicales, alors même qu'il n'est pas envisageable d'obtenir des données annotées suffisamment complètes pour entraîner un modèle supervisé efficace. Les terminologies sont alors un moyen de supervision distante potentiellement efficace (Lerner *et al.*, 2020).

Nous avons ainsi participé à la mise en œuvre du processus de désidentification utilisé dans l'entrepôt de données de santé de l'AP-HP, avec une approche hybride permettant de donner le meilleur des données structurées disponibles, de règles et d'un modèle d'apprentissage profond (Paris *et al.*, 2019). Seule cette hybridation a abouti à des résultats permettant la mise à disposition de documents pseudonymisés pour la recherche.

D'autre part, des collaborations entre informaticiens et médecins conduisent à la définition d'un problème précis de caractérisation de patient (phénotypage), comme par exemple la détection des patients répondant à des critères d'inclusion ou d'exclusion issus de projets de recherche médicale particuliers. Ces problèmes sont si spécifiques qu'il est impossible d'adopter une approche générique pour les résoudre, mais un protocole de travail et de transmission de l'information entre les différentes spécialistes est en cours d'élaboration, pour faciliter et accélérer la réalisation des outils dans le futur.

8 Discussion et conclusion

Pour commencer, notons que la médecine ne supporte pas l'erreur ou l'approximation. La recommandation de vacances ou d'achat n'a pas les mêmes enjeux que la santé des gens. Les approches sémantiques sont plus longues que les autres mais utiles dans les domaines où les données ne sont pas nombreuses. Les approches numériques sont plus efficaces mais doivent être nourries de données nombreuses et validées.

Les difficultés que rencontrent les 2 approches sont les mêmes, liées à la forte hétérogénéité des textes médicaux. Dans le cadre des approches d'apprentissage, cette hétérogénéité se traduit par des difficultés à expliciter les bons paramètres des modèles. Dans le cadre des modèles symboliques, cela se traduit par des difficultés à mettre en place les bons patrons de repérage syntaxique d'un certain nombre de formes textuelles et par l'association des bons termes et des bons synonymes aux concepts modélisés pour réussir leur repérage dans les textes. Une autre difficulté partagée par les 2 approches est la question des corpus de texte en français pour entraîner et tester les systèmes. C'est un problème spécialement ennuyeux pour les approches par apprentissage.

Enfin, *last but not least*, la qualité des données, ici des corpus, est un problème majeur, ou que les documents sont spécialement mal écrits et de nombreuses procédures de correction, de disambiguïsation doivent être mises en œuvre, ou que le contexte d'élaboration des documents est inconnu ou différent de celui de l'utilisation projetée et les corpus et les données peuvent être néfastes à la mise au point du système d'IA (GIGO [6]).

Finalement, plusieurs constats et conclusions s'imposent. D'une part, les approches basées uniquement sur l'annotation massive de données comme unique medium de transfert de la connaissance, qui sont devenues la norme en reconnaissance d'images par exemple,

6. *Garbage in, garbage out.*

n'ont pas fait leurs preuves dans le domaine de l'analyse des textes cliniques. D'autre part, tout outil automatique, même plus performant que l'humain (ce qui n'est pas le cas à l'heure actuelle pour le texte) ne pourra être accepté par les cliniciens et les patients qu'avec une intervention ou une validation humaine, ce qui implique que cet outil doit produire une explication lisible de ses prédictions. Enfin, le nœud du problème se situe au niveau du transfert de connaissances entre l'humain et la machine, transfert qui nécessite d'une part des progrès méthodologiques mais également une collaboration approfondie entre experts de disciplines différentes.

De la même façon que les biostatisticiens et les bio-informaticiens ont intégré les laboratoires de recherche médicale il y a quelques décennies, une discipline et un métier nouveaux doivent être créés, qui permettent de dépasser les collaborations interdisciplinaires actuelles et d'intégrer réellement la science des données et des connaissances dans les services. Fortes de ce constat, des filières de formation s'organisent partout en France dans cette optique ; aux institutions de suivre ce mouvement pour créer des postes permettant de déclencher réellement la révolution attendue dans la santé. On peut finalement reprendre l'affirmation des Dr Solert et Bondu [7] :

> *Et paradoxalement, c'est justement peut-être parce ce qu'aujourd'hui « l'intelligence Artificielle » est d'un niveau extrêmement faible, et qu'elle ne peut toujours pas s'opposer aux décisions du médecin. Elle est toujours un appoint cognitif, un accélérateur pour les choix à faire et des décisions à prendre, ces dernières n'étant toujours que du ressort de l'équipe médicale restreinte entourant le Patient.*

Et c'est probablement dans ce contexte qu'il faut faire progresser les systèmes d'IA.

Références

AGARWAL V., PODCHIYSKA T., BANDA J. M., GOEL V., LEUNG T. I., MINTY E. P., SWEENEY T. E., GYANG E. & SHAH N. H. (2016). Learning statistical models of phenotypes using noisy labeled training data. *J Am Med Inform Assoc*, **23**, 1166–1173.

BEAULIEU-JONES B. K. & GREENE C. S. (2016). Semi-supervised learning of the electronic health record for phenotype stratification. *Journal of Biomedical Informatics*, **64**, 168–178.

CARDOSO S. (2019). *Apports de la modélisation ontologique pour l'analyse des ruptures de parcours de soins dans la Sclérose Latérale Amyotrophique*. phdthesis, Sorbonne Université. Accessible à https://tel.archives-ouvertes.fr/tel-02429414.

CARROLL R. J., EYLER A. E. & DENNY J. C. (2015). Intelligent Use and Clinical Benefits of Electronic Health Records in Rheumatoid Arthritis. *Expert review of clinical immunology*, **11**(3), 329–337.

DENNY J. C., BASTARACHE L., RITCHIE M. D., CARROLL R. J., ZINK R., MOSLEY J. D., FIELD J. R., PULLEY J. M., RAMIREZ A. H., BOWTON E., BASFORD M. A., CARRELL D. S., PEISSIG P. L., KHO A. N., PACHECO J. A., RASMUSSEN L. V., CROSSLIN D. R., CRANE P. K., PATHAK J., BIELINSKI S. J., PENDERGRASS S. A., XU H., HINDORFF L. A., LI R., MANOLIO T. A., CHUTE C. G., CHISHOLM R. L., LARSON E. B., JARVIK G. P., BRILLIANT M. H., MCCARTY C. A., KULLO I. J., HAINES J. L., CRAWFORD D. C., MASYS D. R. & RODEN D. M. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature biotechnology*, **31**(12), 1102–1110.

DYNOMANT E., LELONG R., DAHAMNA B., MASSONNAUD C., KERDELHUÉ G., GROSJEAN J., CANU S. & DARMONI S. J. (2019). Word embedding for the french natural language in health care : Comparative study. *JMIR Med Inform*, **7**(3), e12310.

GOTTESMAN O., KUIVANIEMI H., TROMP G., FAUCETT W. A., LI R., MANOLIO T. A., SANDERSON S. C., KANNRY J., ZINBERG R., BASFORD M. A., BRILLIANT M., CAREY D. J., CHISHOLM R. L., CHUTE C. G., CONNOLLY J. J., CROSSLIN D., DENNY J. C., GALLEGO C. J., HAINES J. L., HAKONARSON H., HARLEY J., JARVIK G. P., KOHANE I., KULLO I. J., LARSON E. B., MCCARTY C., RITCHIE M. D., RODEN D. M., SMITH M. E., BÖTTINGER E. P., WILLIAMS M. S., & EMERGE NETWORK T. (2013). The Electronic Medical Records and Genomics (eMERGE) Network : past, present, and future. *Genetics in Medicine*, **15**(10), 761–771.

7. https://innovationesante.fr/linformatique-cognitive-de-watson-dibm-au-service-de-lhomme-watson-health-espoirs-et-limites/

HALPERN Y., HORNG S., CHOI Y. & SONTAG D. (2016). Electronic medical record phenotyping using the anchor and learn framework. *J Am Med Inform Assoc*, **23**, 731–740.

LEE W.-S., AHN S. M., CHUNG J.-W., KWON K. O. K. K. A., KIM Y. & SYM S. (2018). Assessing concordance with watson for oncology, a cognitive computing decision support system for colon cancer treatment in korea. *JCO Clinical Cancer Informatics*, p. 1–8.

LERNER I., PARIS N. & TANNIER X. (2020). Terminologies augmented recurrent neural network model for clinical named entity recognition. *Journal of Biomedical Informatics*, **102**.

LIN C., KARLSON E. W., DLIGACH D., RAMIREZ M. P., MILLER T. A., MO H., BRAGGS N. S., CAGAN A., GAINER V., DENNY J. C. & SAVOVA G. K. (2015). Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. *Journal of the American Medical Informatics Association : JAMIA*, **22**(e1), e151–e161.

MADIGAN D., RYAN P. B., SCHUEMIE M., STANG P. E., OVERHAGE J. M., HARTZEMA A. G., SUCHARD M. A., DUMOUCHEL W. & BERLIN J. A. (2013). Evaluating the impact of database heterogeneity on observational study results. *American journal of epidemiology*, **178**, 645–651.

NZALI M. D. T., NÉVÉOL A. & TANNIER X. (2015). Analyse d'expressions temporelles dans les dossiers électroniques patients. In *Actes de la Conférence Traitement Automatique des Langues Naturelles (TALN 2015)*, Caen, France.

NÉVÉOL A., DALIANIS H., VELUPILLAI S., SAVOVA G. & ZWEIGENBAUM P. (2018). Clinical Natural Language Processing in languages other than English : opportunities and challenges. *Journal of Biomedical Semantics*, **9**, 12.

PARIS N., DOUTRELIGNE M., PARROT A. & TANNIER X. (2019). Désidentification de comptes-rendus hospitaliers dans une base de données OMOP. In *Actes de TALMED 2019 : Symposium satellite francophone sur le traitement automatique des langues dans le domaine biomédical*, Lyon, France.

RITCHIE M. D., VERMA S. S., HALL M. A., GOODLOE R. J., BERG R. L., CARRELL D. S., CARLSON C. S., CHEN L., CROSSLIN D. R., DENNY J. C., JARVIK G., LI R., LINNEMAN J. G., PATHAK J., PEISSIG P., RASMUSSEN L. V., RAMIREZ A. H., WANG X., WILKE R. A., WOLF W. A., TORSTENSON E. S., TURNER S. D. & MCCARTY C. A. (2014). Electronic medical records and genomics (eMERGE) network exploration in cataract : Several new potential susceptibility loci. *Molecular Vision*, **20**, 1281–1295.

ROSENBLOOM S. T., MILLER R. A. & JOHNSON K. B. (2006). Interface terminologies : facilitating direct entry of clinical data into electronic health record systems. *J Am Med Inform Assoc*, **13**(3), 277–88.

SIMON G., DINARDO C. D., TAKAHASHI K., CASCONE T., POWERS C., STEVENS R. & ALLEN J. (2019). Applying artificial intelligence to address the knowledge gaps in cancer care. *The Oncologist*, **24**(6), 772-82.

STRÖTGEN J. & GERTZ M. (2013). Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*, **47**(2), 269–298.

TAPI NZALI M. D., NÉVÉOL A. & TANNIER X. (2015). Automatic Extraction of Time Expressions Accross Domains in French Narratives. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2015, short paper)*, Lisbon, Portugal.

TOURILLE J., FERRET O., TANNIER X. & NÉVÉOL A. (2017). Neural Architecture for Temporal Relation Extraction : A Bi-LSTM Approach for Detecting Narrative Containers. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017, short paper)*, Vancouver, Canada.

TVARDIK N., KERGOURLAY I., BITTAR A., SEGOND F., DARMONI S. & METZGER M.-H. (2018). Accuracy of using natural language processing methods for identifying healthcare-associated infections. *International Journal of Medical Informatics*, **117**, 96–102.

WEI W.-Q. & DENNY J. C. (2015). Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Medicine*, **7**(1), 41.

WU S., MILLER T., MASANZ J., COARR M., HALGRIM S., CARRELL D. & CLARK C. (2014). Negation's not solved : generalizability versus optimizability in clinical natural language processing. *PloS One*, **9**(11), e112774.

XU Y., ZHOU X., SUEHS B. T., HARTZEMA A. G., KAHN M. G., MORIDE Y., SAUER B. C., LIU Q., MOLL K., PASQUALE M. K., NAIR V. P. & BATE A. (2015). A Comparative Assessment of Observational Medical Outcomes Partnership and Mini-Sentinel Common Data Models and Analytics : Implications for Active Drug Safety Surveillance. *Drug safety*, **38**, 749–765.