AfIA

Association française
pour l'Intelligence Artificielle

Actes de la journée Santé & IA

29 juin 2020

Avec le soutien de l'Association française d'Informatique Médicale (AIM) et le Collège Science de l'Ingénierie des Connaissances de l'AFIA

Dans le cadre de la
Plate-Forme Intelligence Artificielle (PFIA)

PFIA 2020
ANGERS

Table des matières

Comités

Responsables de la journée

Fleur Mougin, ERIAS, Université de Bordeaux & INSERM, Bordeaux
Lina Soualmia, LITIS & LIMICS, Normandie Universités, Rouen

Comité d'initiative

Sandra Bringay, LIRMM, Université Montpellier-3 & CNRS, Montpellier
Jean Charlet, LIMICS, INSERM & AP-HP, Paris
Brigitte Séroussi, LIMICS, Sorbonne Université & AP-HP, Paris
Lina Soualmia, LITIS & LIMICS, Normandie Universités, Rouen
Nathalie Souf, IRIT, Université Paul Sabatier & CNRS, Toulouse
Lynda Tamine-Lechani, IRIT, Université Paul Sabatier & CNRS, Toulouse

Comité de programme

Sandra Bringay, LIRMM, Université Montpellier-3 & CNRS, Montpellier
Jean Charlet, LIMICS, INSERM & AP-HP, Paris
Adrien Coulet, LORIA, Université de Lorraine & CNRS, Nancy
Marc Cuggia, LTSI, Université de Rennes 1 & INSERM, Rennes
Olivier Dameron, IRISA, Université de Rennes 1 & CNRS, Rennes
Gayo Diallo, ERIAS, Université de Bordeaux & INSERM, Bordeaux
Natalia Grabar, STL, Université de Lille & CNRS, Lille
Clément Jonquet, LIRMM, Université de Montpellier & CNRS, Montpellier
Fleur Mougin, ERIAS, Université de Bordeaux & INSERM, Bordeaux
Lemlih Ouchchane, Université Clermont Auvergne, Clermont-Ferrand
Brigitte Séroussi, LIMICS, Sorbonne Université & INSERM & AP-HP, Paris
Lina Soualmia, LITIS & LIMICS, Normandie Universités & INSERM, Rouen
Nathalie Souf, IRIT, Université Paul Sabatier & CNRS, Toulouse
Lynda Tamine-Lechani, IRIT, Université Paul Sabatier & CNRS, Toulouse
Pierre Zweigenbaum, LIMSI, Université Paris-Saclay & CNRS, Orsay

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.

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.

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 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]).

6. *Garbage in, garbage out.*

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, 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.,

7. Cf. URL de la note 4

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.

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.

Utilisation des graphes pour la représentation spatio-temporelle lors d'un examen d'IRM fonctionnelle cérébrale

Aurélie Leborgne[1], Florence Le Ber[1], David Niezgoda[1], Céline Meillier[1], Stella Marc-Zwecker[1]

UNIVERSITÉ DE STRASBOURG, CNRS, ENGEES, ICUBE UMR 7537,
F-67000 Strasbourg, France
aurelie.leborgne@unistra.fr

Résumé : Les données issues de l'imagerie médicale ont très souvent des caractéristiques à la fois spatiales et temporelles, nécessitant une modélisation spécifique. Nous nous intéressons ici à l'activité cérébrale et nous proposons de la modéliser par un graphe spatio-temporel. Nous traitons des données collectées au cours d'un examen IRMf réalisé sur une population animale, indiquant l'activité conjointe observée entre les aires (zones spatiales) du cerveau. La méthode proposée permet de grouper les aires cérébrales en réseaux sur des fenêtres temporelles, en fonction de la similarité de leurs niveaux d'activation. Les réseaux existants dans une fenêtre temporelle sont ensuite représentés comme les noeuds d'un graphe, et reliés aux réseaux de la fenêtre suivante par des relations topologiques, pour former un graphe spatio-temporel établissant la chronologie des réseaux de co-activation. De plus, des visualisations sont proposées.

Abstract : Medical imaging produces a large volume of data with spatial and temporal characteristics, requiring thus a specific modelling. We focus here on brain activiy and propose to use spatio-temporal graphs. Data were collected during a fMRI examination on small animals, showing the conjoint activity of brain (spatial) areas. The proposed method allows to group brain areas into networks for a time window, according to the similarity of their activity level. Networks of a time window are then represented as the nodes of a graph, and linked to the networks of the following time window. This results in a spatio-temporal graph. representing the chronology of network co-activations. Furthermore, graph visualizations are provided.

Mots-clés : IRM fonctionnelle, graphe spatio-temporel, modélisation des données, relation topologique, classification, visualisation.

1 Introduction

Dans le domaine des neurosciences, les outils et techniques s'améliorent régulièrement (Orrison *et al.*, 2017). Ces évolutions augmentent la précision et le nombre de données recueillies lors des examens cérébraux. De même que ce développement quantitatif, on observe également une amélioration qualitative. En effet, les études du cerveau intègrent aujourd'hui une dimension temporelle, sur des échelles de temps de plus en plus fines. On observe alors une complexification du traitement, de l'analyse et de l'interprétation de ces informations. Ceci accroît la difficulté à élaborer des modèles explicatifs des faits observés, intégrant les dimensions spatiales mais également les aspects chronologiques (Atluri *et al.*, 2018).

L'objectif des recherches en neuro-imagerie consiste principalement à déterminer s'il existe des motifs d'activité au sein du cerveau (Leonardi & Van De Ville, 2015; Vidaurre *et al.*, 2017; Cabral *et al.*, 2017). Plus spécifiquement, on cherche à mettre en évidence les séquences d'évènements que forment les activations mutuelles des aires cérébrales dans le temps. Ceci permet d'observer s'il existe, notamment dans le cadre d'études sur les troubles mentaux, des perturbations ou des irrégularités par rapport aux activations normales des aires cérébrales, au regard de ces motifs (Atluri *et al.*, 2018; Sourty, 2016; Damaraju *et al.*, 2014).

Certaines études suggèrent d'analyser ces relations spatio-temporelles en les représentant sous la forme de graphes (Atluri *et al.*, 2018; Sourty, 2016). Il est également possible de les représenter sous la forme de graphes spatio-temporels (Del Mondo *et al.*, 2013), c'est ce que nous verrons dans la suite de cet article. Longtemps utilisées en sciences de l'information géographique, ces méthodes s'appliquent également à des domaines comme la santé ou les neurosciences (Atluri *et al.*, 2018). Ces approches permettent de modéliser des entités et les

relations qui les lient, autant de manière topologique que chronologique. Des opérations algébriques peuvent également être appliquées, comme le propose le modèle *Region Connection Calculus* (Randell *et al.*, 1992), notamment dans sa variante à 5 relations (RCC5).

L'objectif de l'étude présentée dans cet article est de développer une méthode de modélisation originale adaptée au traitement de données issues d'imagerie cérébrale et permettant de visualiser la chronologie des activations des réseaux reliant les différentes aires cérébrales. En section 2, nous détaillerons les données utilisées, qui proviennent d'examens d'IRM fonctionnelle pratiqués sur des populations de souris. En section 3 nous présentons notre approche, inspirée des travaux de la littérature (Sourty, 2016; Del Mondo *et al.*, 2013). La section 4 suivante présente des résultats au moyen de deux types de visualisations, l'une spatiale (visualisation des réseaux et des aires à un temps donné), l'autre temporelle (visualisation des relations entre réseaux dans des temps successifs). Puis nous concluons (section 5).

2 Acquisition des données

2.1 Définitions

L'étude de la connectivité fonctionnelle cérébrale nécessite, dans un premier temps, de délimiter les régions d'intérêt dans le cerveau puis, dans un second temps, de définir les relations qui lient ces différentes régions d'intérêt. Des segmentations anatomiques du cerveau de la souris existent à différents niveaux. Dans ce papier, nous parlerons de régions pour la segmentation anatomique la plus grossière. Cependant nous pouvons distinguer une ségrégation fonctionnelle à l'intérieur de ces grandes régions anatomiques. Une segmentation anatomique plus fine permet de séparer les régions en un ensemble d'aires cérébrales. Au sein d'une même région, les aires ne présentent pas forcément une activité temporelle identique, puisque par définition elles ne répondent pas forcément à la même fonction malgré leur proximité spatiale. Plusieurs aires d'une même région peuvent avoir une activité similaire et sont regroupées dans un réseau, et, au cours du temps, une même aire peut faire partie de plusieurs réseaux au sein de la région. Des réseaux de différentes régions peuvent ensuite présenter des co-activations afin de répondre à une même fonction, on parlera dans ce cas de réseau fonctionnel.

Une représentation graphique de ces différentes notions est donnée dans la figure 1. Sur cette figure, chaque région a une couleur qui lui est propre (Région 1 en bleu, Région 2 en vert et Région 3 en jaune). Chacune des régions est constituée d'aires (représentées par des disques colorés). Les réseaux, regroupant des aires, sont entourés en fuschia. De plus, les réseaux 1 et 3 ainsi que les réseaux 2 et 4 ont une activité corrélée. Cette co-activation entre réseaux est représentée par une ligne fuschia. Les réseaux 1 et 3 d'une part, ainsi que les réseaux 2 et 4 d'autre part, constituent ainsi des réseaux fonctionnels.

2.2 Acquisition et pré-traitements

Les données sont acquises lors d'un examen pré-clinique réalisé sur deux groupes de souris dans le but d'étudier l'évolution de la maladie d'Alzheimer et sa progression dans le cerveau. L'objectif, à terme, est de comprendre l'impact de cette maladie en termes de connectivité entre les réseaux neuronaux.

Les données sont obtenues au moyen d'une IRM fonctionnelle à haut champ de 7 teslas. Cet outil permet d'observer localement le fonctionnement du cerveau, de manière indirecte, en enregistrant les variations des propriétés du flux sanguin (Rodden & Stemmer, 2008). A l'issue d'un examen réalisé sur un animal, une valeur correspondant au signal qui reflète l'activité neuronale dans la zone cérébrale associée est obtenue pour chaque voxel (élément de l'image en 3 dimensions).

Un pré-traitement des données est nécessaire afin de réaligner les images par rapport aux mouvements du sujet, de les normaliser et de lisser les différences individuelles dans l'objectif de réaliser des études de groupe. Le recalage des données sur un atlas auquel est associée une carte de segmentation du cerveau, où chaque classe correspond à une aire cérébrale, permet ensuite d'estimer le signal temporel associé à chaque aire pour chacune des souris. Pour une

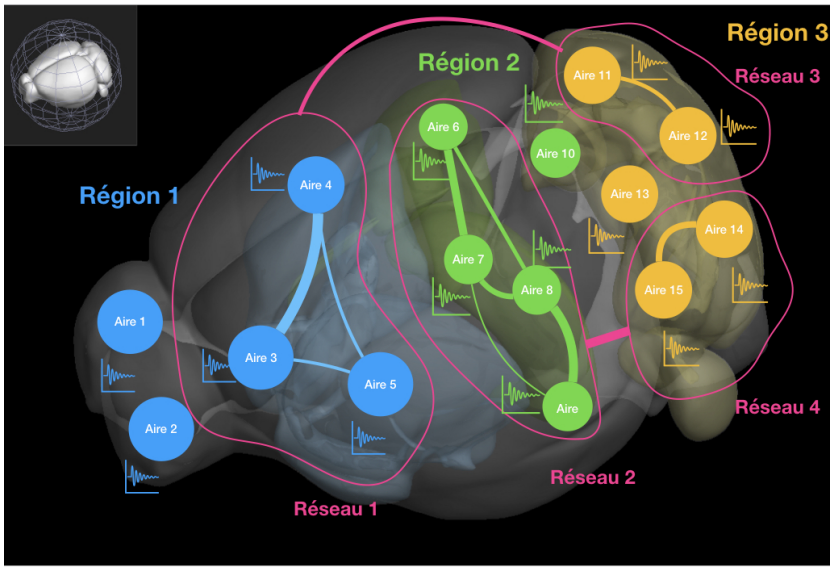


FIGURE *1 – Représentation schématique des notions de région, d'aire et de réseau, sur la base d'une capture d'écran du logiciel Brain Explorer développé par le Allen Institute*

souris donnée, il est alors possible de mesurer la ressemblance des signaux temporels des différentes aires cérébrales *via*, par exemple, le calcul d'une matrice de corrélation comme l'illustre la figure 2. Plus deux aires sont fortement corrélées (*resp.* anti-corrélées) et plus le carré correspondant sur la matrice de corrélation est rouge (*resp.* bleu).

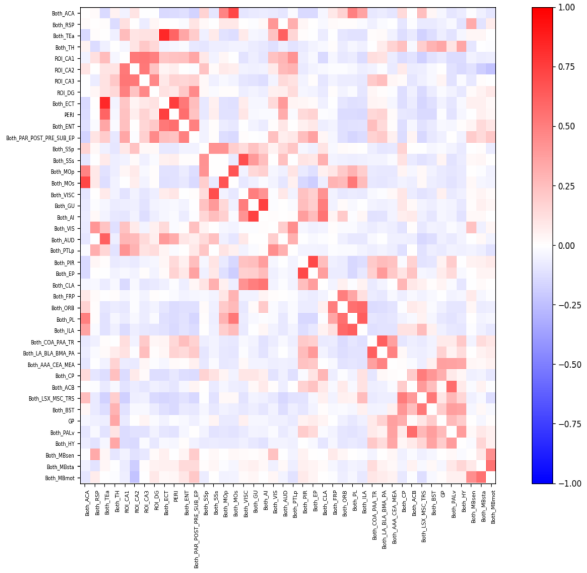


FIGURE *2 – Représentation graphique d'une matrice de corrélation à un temps donné*

Dans le cadre de l'étude de la dynamique de la connectivité cérébrale, ce sont les relations entre les aires sur des petites fenêtres temporelles successives qui nous intéressent. Nos données d'entrée correspondent alors à une série de matrices de corrélation. Elles indiquent,

pour chaque temps, les taux d'activation conjointe de chaque couple d'aires cérébrales, représentées en ligne et en colonne. Pour chaque matrice M^t, associée à une fenêtre temporelle t, la valeur $M^t(x, y)$ correspond à la corrélation temporelle des signaux des deux aires x et y sur la fenêtre t. L'utilisation d'un seuil s_R permet d'éliminer les corrélations les plus basses (en valeur absolue) et de déterminer ainsi des sous-ensembles d'aires reliées par une forte corrélation ou anti-corrélation, qui forment ainsi des réseaux.

3 Modélisation des données

3.1 Un modèle de graphe spatio-temporel

Dans la suite, on considère un ensemble de n matrices de corrélation seuillées, ordonnées dans le temps. Chaque matrice M^{t_i}, $i \in [1, n]$, est représentée par un graphe de corrélation, noté $\mathcal{G}^{t_i}$, dont les sommets sont les réseaux existant au temps t_i, et les arêtes portent les corrélations calculées entre ces réseaux ; ces graphes sont ensuite reliés entre eux deux-à-deux, $\mathcal{G}^{t_i}$ et $\mathcal{G}^{t_{i+1}}$, $i \in [1, n-1]$, par des arêtes représentant les relations spatiales entre réseaux de deux temps successifs. Nœuds et arêtes sont étiquetés.

L'ensemble ainsi constitué est un graphe spatio-temporel (graphe-ST), que nous formalisons de la façon suivante, sur la base de la proposition de (Del Mondo *et al.*, 2013). Nous introduisons un domaine temporel, $\mathcal{T} = \{t_1, t_2, ...t_n\}$, où t_i représente une instance de temps d'une granularité donnée et $t_i < t_{i+1}$ pour tout $i \in [1, n-1]$. Δ est un ensemble d'entités, $\{e_1, e_2, ...e_m\}$. Nous introduisons également Ξ, un ensemble de relations de corrélation et Σ, un ensemble de relations spatio-temporelles.

Un graphe-ST $\mathcal{G}$ est un tuple (V, E_Ξ, E_Σ, L), où V est un ensemble de sommets $(e_l, t_i) \in \Delta \times \mathcal{T}$, E_Ξ est un ensemble de tuples $((e_i, t_i)\xi(e_j, t_i))$ où $(e_i, t_i), (e_j, t_i) \in V$ et $\xi \in \Xi$. E_Σ est un ensemble de tuples $((e_i, t_i)\sigma(e_j, t_{i+1}))$ où $(e_i, t_i), (e_j, t_{i+1}) \in V$ et $\sigma \in \Sigma$. Finalement $L = L_\Delta \cup L_\Xi$ est un ensemble d'étiquettes, qui portent respectivement sur les ensembles Δ et Ξ. Nous décrivons ci-dessous les différents éléments d'un graphe spatio-temporel.

Nœuds

Un nœud (e_l, t_i) représente un réseau $e_l \in \Delta$ existant au temps $t_i \in \mathcal{T}$. Comme nous l'avons vu dans la section 2, un réseau est composé d'aires du cerveau qui appartiennent à une même région et sont fortement liées entre elles (liaison mesurée par rapport à un seuil s_R) de manière directe ou indirecte. De plus, chaque réseau porte une étiquette $l_e \in L$ qui représente ici la valeur maximale des corrélations entre ses aires, valeur qui constitue une information sur la connectivité interne du réseau ($|l_e| \in [s_R, 1]$). Les aires ne sont donc pas représentées directement dans le graphe, mais via les réseaux dont elles font partie.

Arêtes de corrélation

Les nœuds d'un graphe $\mathcal{G}^{t_i}$ sont reliés par des arêtes de corrélation $\xi \in \Xi$. Ces arêtes sont établies à partir des valeurs de corrélation entre les aires constituant les réseaux. Ainsi la valeur de corrélation entre deux réseaux e_l, e_m est égale à la valeur maximale de corrélation entre chacune des aires de e_l et chacune des aires de e_m. Si cette valeur est supérieure ou égale à un seuil donné s_Ξ, alors une arête $((e_l, t_i)\xi(e_m, t_i))$ est introduite dans le graphe $\mathcal{G}^{t_i}$, où ξ porte une étiquette $l_\xi \in L$ avec $|l_\xi| \in [s_\Xi, 1]$. L'utilisation d'un seuil a pour but de faire abstraction du bruit.

Arêtes spatio-temporelles

Les réseaux évoluent dans le temps, c'est-à-dire que d'un temps t_i à un temps t_{i+1} les aires n'appartiennent pas forcément aux mêmes réseaux. Pour modéliser cette dynamique, on utilise le modèle RCC5 de relations topologiques (Randell *et al.*, 1992). Considérons deux réseaux e_l et $e_{l'}$ rattachés à deux sous-graphes successifs $\mathcal{G}^{t_i}$ et $\mathcal{G}^{t_{i+1}}$ (*cf.* figure 3) :

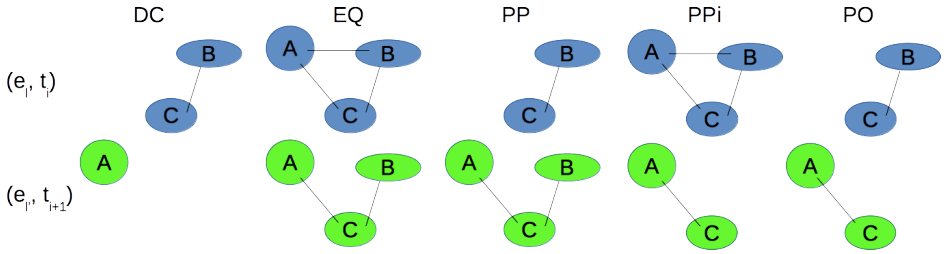


FIGURE 3 – *Les différentes configurations de deux réseaux à deux temps successifs : A, B, C représentent les aires à l'intérieur des réseaux e_l (présent au temps t_i) et $e_{l'}$ (présent au temps t_{i+1})*

— si les réseaux ne se recouvrent pas du tout (pas d'aire commune), alors il y a déconnexion (DC) entre les deux réseaux, ce qui se traduit par l'absence d'arête ;

— si les réseaux couvrent exactement le même ensemble d'aires, alors il y a égalité spatiale (EQ) entre les deux réseaux, ce qui se traduit par une arête $((e_l, t_i)\texttt{EQ}(e_{l'}, t_{i+1}))$;

— si toutes les aires de e_l appartiennent aussi à $e_{l'}$ (mais la réciproque n'est pas vraie), alors e_l est partie propre (PP) de $e_{l'}$, ce qui se traduit par une arête $((e_l, t_i)\texttt{PP}(e_{l'}, t_{i+1}))$;

— si inversement, toutes les aires de $e_{l'}$ appartiennent à e_l (mais la réciproque n'est pas vraie), alors e_l contient (PPi) $e_{l'}$, ce qui se traduit par une arête $((e_l, t_i)\texttt{PPi}(e_{l'}, t_{i+1}))$;

— si les réseaux e_l et $e_{l'}$ ont des aires communes et des aires propres, alors e_l et $e_{l'}$ se recouvrent partiellement (PO), ce qui se traduit par une arête $((e_l, t_i)\texttt{PO}(e_{l'}, t_{i+1}))$.

Finalement l'ensemble des relations spatio-temporelles est $\Sigma = \{\texttt{EQ}, \texttt{PP}, \texttt{PPi}, \texttt{PO}\}$. La déconnexion DC est représentée implicitement par l'absence d'arête.

3.2 Exemple

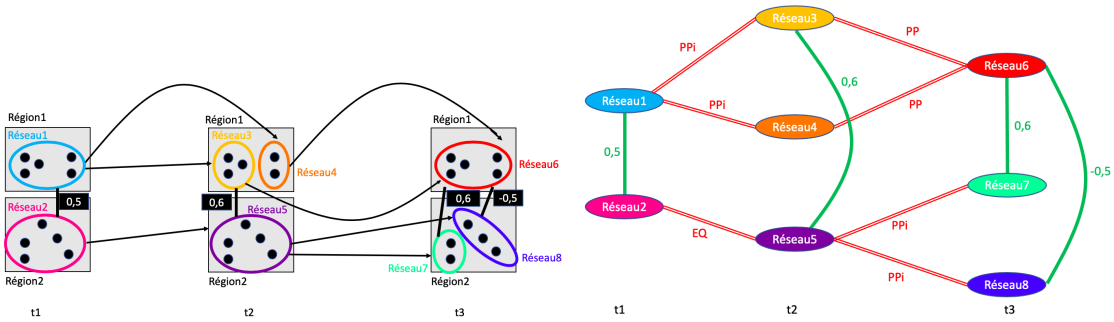
La figure 4(a) représente l'évolution de huit réseaux à trois instants successifs d'un examen d'IRM fonctionnelle. Cette évolution est représentée par un graphe-ST $\mathcal{G}_1$ (figure 4(b)) construit sur les domaines $\mathcal{T} = \{t_1, t_2, t_3\}$ et $\Delta = \{\texttt{Réseau}_1, \texttt{Réseau}_2, \texttt{Réseau}_3, \texttt{Réseau}_4, \texttt{Réseau}_5, \texttt{Réseau}_6, \texttt{Réseau}_7, \texttt{Réseau}_8\}$. De plus, les relations de corrélation sont étiquetées par des valeurs discrètes de l'intervalle $[-1; -0,4] \cup [0,4; 1]$ et les relations spatio-temporelles proviennent de la théorie RCC5 (Randell *et al.*, 1992), comme expliqué ci-dessus.

— Le sous-graphe des relations de corrélation $\mathcal{G}_1^{t_i}$ représente les corrélations existant entre les réseaux à un temps donné $t_i \in \mathcal{T}$. Sur la figure 4(b), ces relations sont représentées par des lignes vertes. Par exemple, les nœuds $(\texttt{Réseau}_3, t_2)$ et $(\texttt{Réseau}_5, t_2)$ sont connectés par une arête étiquetée avec une valeur de corrélation de $0,6$.

— Le sous-graphe des relations spatio-temporelles représente les interactions spatiales entre réseaux à deux instances de temps successives. Dans la figure 4(b), ces relations sont représentées par une double ligne rouge. Par exemple, les nœuds $(\texttt{Réseau}_1, t_1)$ et $(\texttt{Réseau}_3, t_2)$ sont reliés par une arête PPi alors que les nœuds $(\texttt{Réseau}_2, t_1)$ et $(\texttt{Réseau}_5, t_2)$ sont connectés par une arête EQ.

4 Résultats

Les graphes spatio-temporels construits sont complexes et donc difficilement visualisables. C'est pourquoi nous proposons une double visualisation, qui met en exergue, d'une part, les corrélations spatiales des réseaux à un temps donné, et, d'autre part, les réorganisations temporelles des réseaux.



(a) Données sur les réseaux au cours du temps (b) Le graphe correspondant $\mathcal{G}_1$

FIGURE *4 – Un exemple de graphe construit à partir de données IRMf*

4.1 Visualisation spatiale

À partir d'un graphe-ST issu d'un ensemble de matrices, nous proposons une visualisation permettant de mettre en évidence les différentes régions, les réseaux, les aires qui les composent et les réseaux fonctionnels, pour chaque fenêtre temporelle. Sur la figure 5, ces différents éléments sont représentés dans une forme circulaire, avec une échelle de couleur pour les valeurs de corrélation. Les régions se situent à l'extérieur du cercle. En allant vers l'intérieur du cercle, chaque aire est reliée à la région à laquelle elle appartient. Puis, en regardant encore plus vers l'intérieur du cercle, les petits disques de couleur blanche à rouge représentent les réseaux à l'instance de temps t_i. Plus la couleur d'un disque se rapproche du rouge et plus les signaux temporels des aires le constituant sont fortement corrélés. À titre d'exemple, le réseau constitué des aires both_MOs et Both_MOp (du cortex sensoriel, encadrées en violet) est un réseau fortement corrélé. De même, l'ensemble des aires du cortex pré-frontal (encadrées en bleu) forme un réseau corrélé. Les segments au centre du cercle dont les couleurs varient du bleu au rouge indiquent la co-activation entre deux réseaux : deux réseaux reliés par un segment bleu ont des activités temporelles anti-corrélées, au contraire deux réseaux reliés par un segment rouge foncé ont une activité temporelle fortement corrélées. Ainsi, les deux réseaux cités ci-dessus sont co-activés (ils sont reliés par une arête rouge foncé, mise en évidence par la flèche rouge), ils font donc partie du même réseau fonctionnel.

4.2 Visualisation temporelle

Un graphe spatio-temporel a été construit pour une séquence d'environ 500 images IRMf, analysées par une quarantaine de petites fenêtres temporelles glissantes, et traduites en autant de matrices de corrélations. La dimension temporelle du graphe résultant peut être visualisée comme le montre la figure 6. Les relations spatio-temporelles entre réseaux de deux temps successifs sont dénotées par différentes couleurs, par exemple le rouge pour la relation spatio-temporelle partie-de (PP). Cette visualisation met en évidence les évolutions des réseaux d'aires au cours du temps et à l'intérieur des différentes régions du cerveau.

On remarque ainsi différents comportements des régions. Par exemple, les aires qui constituent l'insula et le cortex préfrontal forment un ou plusieurs réseaux qui ne présentent pas ou peu de réorganisation ; au contraire, dans le cortex sensoriel, le taux de réorganisation des aires en réseaux est très élevé. Plus précisément, on peut interpréter le comportement de l'insula de la façon suivante : au départ les deux aires qui la constituent ont une activité différente, puis les deux aires se synchronisent en un seul réseau : elles ont alors une activité qui peut varier, mais toujours de façon corrélée. Dans le cortex sensoriel, l'activité des aires varie souvent, avec des corrélations momentanées pour différentes parties d'entre elles. Pour compléter cette analyse, il faut revenir à l'image présentée en figure 5 pour les temps considérés.

De plus, cette lecture nécessite évidemment de se référer au seuil de corrélation retenu pour constituer les réseaux ($s_R = 0,4$ dans cet exemple).

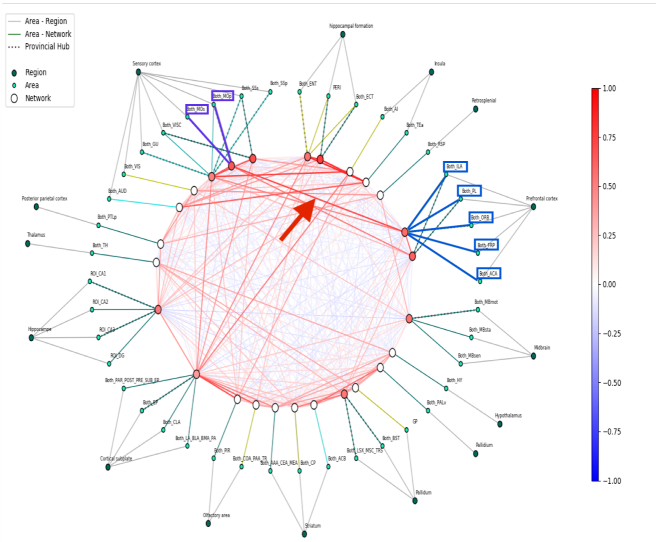


FIGURE 5 – *Visualisation des corrélations à l'intérieur et entre les réseaux à une instance de temps* t_i

5 Conclusion et travaux à venir

Dans cet article, nous avons présenté une proposition pour modéliser les données spatio-temporelles issues d'un examen d'IRMf par un graphe-ST. Ce dernier permet, d'une part, de représenter les relations de corrélation entre les différents réseaux sur chacunes des fenêtres temporelles au cours d'un examen d'IRMf. D'autre part, il permet de modéliser l'évolution des réseaux pendant toute la durée de l'examen. Nous avons également présenté deux propositions pour visualiser ces données dans leurs dimensions spatiale et temporelle.

Cette approche a été expérimentée sur d'autres types de données, comme l'évolution des parcelles agricoles (Leborgne *et al.*, 2019). Le modèle théorique est semblable, cependant les types de relation sont différents. Nous n'avons pas utilisé des relations de corrélation mais des relations spatiales qualitatives (voisinages spatiaux). Les problèmes sur ce type de données sont différents : il y a moins de relations entre les différents nœuds, moins d'instances de temps, mais il y a beaucoup plus d'objets.

Après ce travail de modélisation, il sera intéressant d'analyser et de synthétiser le graphe spatio-temporel ainsi obtenu de manière à mettre en avant les évènements principaux, survenus au cours de l'ensemble de l'examen cérébral. Ces évènements correspondent aux périodes de temps où de nombreuses aires cérébrales se sont activées conjointement, avec une intensité suffisamment forte (forte corrélation ou anti-corrélation). Ces analyse et synthèse de graphe pourraient également permettre de mettre en évidence les principaux motifs d'activité cérébrale, apparus au cours du temps, et correspondant à la présence répétée de certains sous-graphes. Elles permettraient ainsi de mieux comprendre la manière dont s'organisent fonctionnellement les aires du cerveau par rapport à une tâche donnée.

Remerciements

Nous adressons nos remerciements à Laura Harsan de l'équipe Imagerie Multimodale Intégrative en Santé du laboratoire ICube, qui a fourni les données sur lesquelles nous avons travaillé. Ce travail a été partiellement financé par le laboratoire ICube et l'université de Strasbourg (API et Idex METEC-graphe).

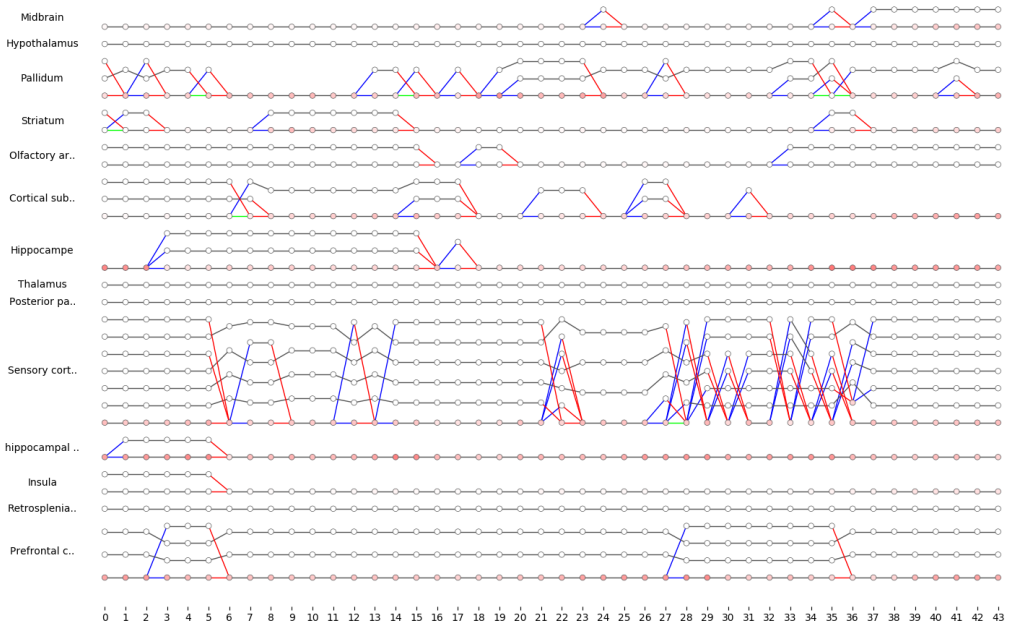


FIGURE 6 – *Evolution des réseaux au cours du temps : en ordonnée sont représentés les réseaux et les régions auquelles ils appartiennent, en abscisse le temps ; les lignes rouges représentent des relations* `PP` *(fusion de réseaux), les lignes bleues des relations* `PPi` *(séparation de réseaux), les lignes vertes des relations* `PO` *(fusion-séparation de réseaux) et les lignes noires des relations* `EQ` *(identité du réseau)*

Références

ATLURI G., KARPATNE A. & KUMAR V. (2018). Spatio-temporal data mining : A survey of problems and methods. *ACM Computing Surveys (CSUR)*, **51**(4), 1–41.

CABRAL J., KRINGELBACH M. L. & DECO G. (2017). Functional connectivity dynamically evolves on multiple time-scales over a static structural connectome : Models and mechanisms. *NeuroImage*, **160**, 84–96.

DAMARAJU E., ALLEN E. A., BELGER A., FORD J. M., MCEWEN S., MATHALON D., MUELLER B., PEARLSON G., POTKIN S., PREDA A. *et al.* (2014). Dynamic functional connectivity analysis reveals transient states of dysconnectivity in schizophrenia. *NeuroImage : Clinical*, **5**, 298–308.

DEL MONDO G., RODRÍGUEZ M. A., CLARAMUNT C., BRAVO L. & THIBAUD R. (2013). Modeling consistency of spatio-temporal graphs. *Data & Knowledge Engineering*, **84**, 59–80.

LEBORGNE A., MEYER A., GIRAUD H., LE BER F. & MARC-ZWECKER S. (2019). Un graphe spatio-temporel pour modéliser l'évolution de parcelles agricoles. In *Actes de la conférence SAGEO, Clermont-Ferrand, France*, p. 1–13.

LEONARDI N. & VAN DE VILLE D. (2015). On spurious and real fluctuations of dynamic functional connectivity during rest. *Neuroimage*, **104**, 430–436.

ORRISON W. W., LEWINE J., SANDERS J. & HARTSHORNE M. F. (2017). *Functional brain imaging*. Elsevier Health Sciences.

RANDELL D. A., CUI Z. & COHN A. G. (1992). A spatial logic based on regions and connection. In *Proceedings 3rd Int Conference on Knowedge Representation and Reasoning*.

RODDEN F. A. & STEMMER B. (2008). A brief introduction to common neuroimaging techniques. In *Handbook of the neuroscience of language*, p. 57–67. Elsevier.

SOURTY M. (2016). *Analyse de la dynamique temporelle et spatiale des réseaux cérébraux spontanés obtenus en imagerie par résonance magnétique fonctionnelle*. PhD thesis, Strasbourg.

VIDAURRE D., SMITH S. M. & WOOLRICH M. W. (2017). Brain network dynamics are hierarchically organized in time. *Proceedings of the National Academy of Sciences*, **114**(48), 12827–12832.

Towards a mobile conversational agent for COVID-19 post quarantine psychological assistance

Nourchène Ouerhani[1], Ahmed Maalel[1,2], Henda Ben Ghézala[1]

[1] University of Manouba, National School of Computer Sciences, RIADI Laboratory, 2010, Manouba, Tunisia
nourchne_ouerhani@yahoo.fr
henda.benghezala@ensi.rnu.tn

[2] University of Sousse, Higher Institute of Applied Sciences and Technology, 4003, Sousse, Tunisia
ahmed.maalel@ensi.rnu.tn

Résumé : Nowadays, several nations in the world is under confinement because of the novel pandemic called COVID-19. This situation has caused feelings of fear, anxiety and depression even suicide. That's why, our purpose is to build a mobile conversational agent for anxiety emotion assistance after COVID-19 quarantine, that communicates with a citizen to increase his/her consciousness towards the real danger of this outbreak. Furthermore, our conversational agent is able to recognize and manage stress mainly after the quarantine period, using natural language understanding (NLU). The messages delivered from our agent and its way of communication could possibly help to avoid anxiety after the world's lockdown. Our proposed approach is a mobile healthcare service that is introduced by its three interdependent units : Input Processing Unit (IPU) in which the natural language understanding (NLU) is done, Storage Unit (SUn) that store every conversation and finally the Response Manager Unit (RMU) that manages the conversational agent answers.

Mots-clés : Conversational Agent, Mobile, Natural Language Understanding, Coronavirus, Mental Health

1 Introduction

The virus that causes COVID-19 is a novel coronavirus that was first identified during an investigation into an outbreak in Wuhan, China [1]. According to Worldometer, on June 03, 2020 more than 380,000 people in world have died as a cause of COVID-19 [2]. Upon the huge proliferation of the pandemic, stress and anxiety are generating throughout the population. People become panicked of this grim. In fact, there is a dramatic difference between panic and awareness. Awareness is the fact of being responsible and conscious of the scale of the problem and things to do. It is normal to be vigilant about COVID-19 and what to do to minimize its spread. But it is better to avoid worsening the situation with negative imagination than fear. Panic is a fear doped with steroids. From the moment the virus entered someone's mental culture, he/she is engulfed in its frightening power.

Given the previously outlined circumstances, a mobile conversational agent for COVID-19 post quarantine psychological assistance according to the considerations presented by the WHO Department of Mental Health and Substance in order to support mental and psychosocial well-being in different target groups during and after the outbreak [3]. The proposed approach is a conversational agent based mobile healthcare service (Silva *et al.*, 2015; Hossain & Muhammad, 2017) that offers healthcare anywhere and to anyone regardless of age, race, gender or socio-demographic status.

The remainder of the paper is organized as follows : first, we take a look at some of the works discussed on COVID-19. Second, we introduce our approach through the functionality of the different units of our conversational agent, in section 3. Then, we present the development progress of our proposed approach's architecture, in 4. And finally, the paper ends with a conclusion 5.

1. https://www.cdc.gov/coronavirus/2019-nCoV/index.html
2. https://www.worldometers.info/coronavirus/
3. https://www.who.int/docs/default-source/coronaviruse/mental-health-considerations.pdf

2 State of the art

Since the first appearance of the COVID-19 in the world, various research works in different domains have focused on this topic. Therefore in the What Is Information Technology (IT) field, several researchers have contributed to the fight against this pandemic. (Alimadadi *et al.*, 2020) shows the impact of Artificial Intelligence and Machine Learning in the world battle against COVID-19. Likewise, (Rao & Vazquez, 2020) proposed to use machine learning techniques to improve the detection of an infected person with COVID-19. Contributions of (Gozes *et al.*, 2020; Barstuğan *et al.*, 2020) belong to the field of medical imaging and aim to build an automated Computed Tomography (CT) image analysis using artificial intelligence and machine learning methods to detect and track COVID-19. In the same field, (Wang & Wong, 2020) presents a deep convolutional neural network called COVID-Net to detect COVID-19 infection from chest X-ray (CXR) images.

According to our research, there is not yet any work which tackles the subject of stress and fear of COVID-19. Even worse, (Goyal *et al.*, 2020) shows that fear of COVID-19 led to suicide so that becomes very dangerous to humanity lives. In this case we decided to offer a conversational agent to reassure people especially those who are infected.

3 Proposed Approach

The architecture is shown in Figure 1. Our adopted approach is introduced by its several units : Input Processing Unit (IPU), Storage Unit (SUn) and Response Manager Unit (RMU).

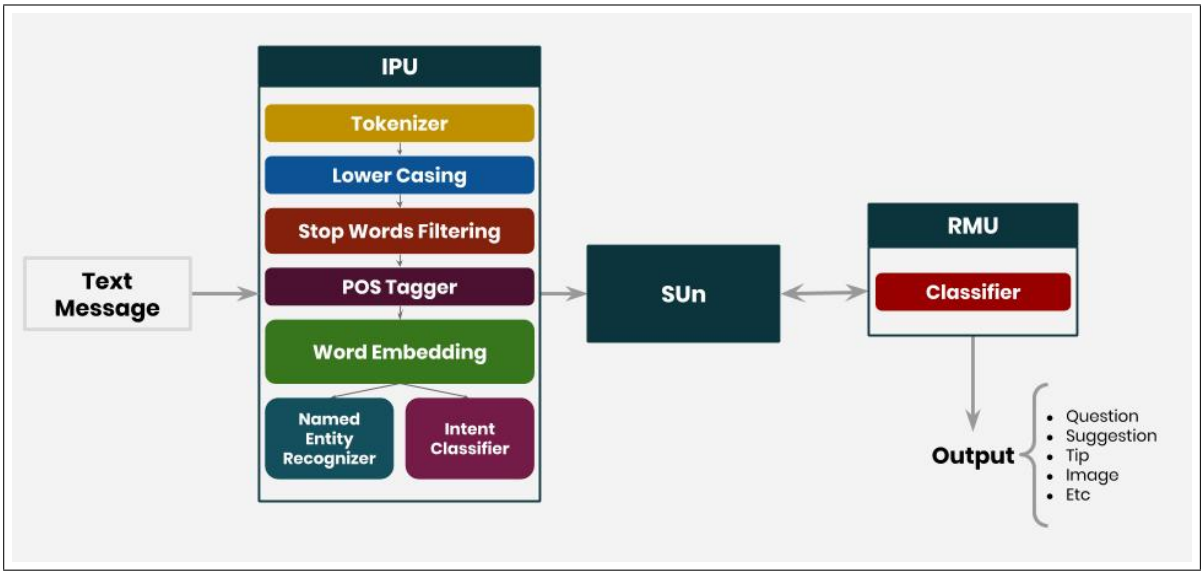


FIGURE *1 – Architecture of the mobile conversational agent*

3.1 Input Processing Unit (IPU) :

Our conversational agent's very first mission is to transform the natural text to a vector representation which is called the natural language processing (NLP), through numerous steps.

3.1.1 Tokenization :

In 1992, (Webster & Kit, 1992) defined tokenization as the first step in NLP that identifies tokens, as shown in Figure 2.

FIGURE *2 – Process of tokenization*

3.1.2 Lower Casing :

After tokenization, we lower case the data as shown in Figure 3

FIGURE *3 – Lower Casing*

3.1.3 Part of Speech tagging (PoS tagging) :

PoS tagging is an important preprocessing task in NLP. According to (Kumar & Josan, 2010), it is a process of giving each word a particular part of speech which can be noun, proper noun, verb, adjective, Determiner, etc as shown in Figure 4.

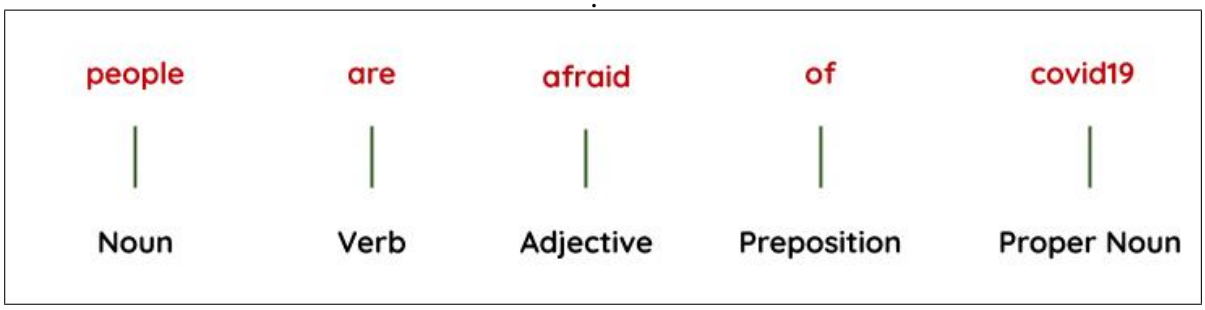


FIGURE *4 – Illustration of PoS tagging*

3.1.4 Stop Word Filtering :

Stop words are frequently used common words. For example, there are a lot of stop words in French such as "au", "car", "ce", "cela", "ces" and "du". These stop words do not contribute to the knowledge source and they can be removed from the textual data. But there is no standard list of stop words. The list of words to ignore can vary depending on the field of the current work.

3.1.5 Word embedding :

The history of NLP is marked by transitions in the ways of representing the input to the model. So, some of the earliest applications (Miikkulainen & Dyer, 1991) attempted to use neural networks for NLP by representing the input as a sequence of characters. And in 2001, the authors of (Bengio *et al.*, 2000) tackled language modeling. At that time, they named this process as “learning a distributed representation for words”. And from here comes the initial

idea of word embedding. Formerly, the traditional one-hot vectors were widely used to represent words. But, this method does not allow capturing the semantic relationship and similarity between words. And thus, it could not be sufficient in NLP tasks. From here comes the word embedding which represents the ideal semantic space of words in a real-valued continuous vector space, and some to capture word similarities as defined in (O'Shea, 2014, 2012). In fact multiple models have appeared such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin *et al.*, 2018) presented in late 2018, fastText (Bojanowski *et al.*, 2016) bearing in mind the famous old ones word2vec (Mikolov *et al.*, 2013) and GloVe (Pennington *et al.*, 2014).

3.1.6 Named Entity Recognition (NER) :

In the expression "Named Entity", the word "Named" aims to restrict the task to only those entities for which one or many rigid designators, as defined by (Kripke, 1980).
Thus, names of people, places, groups, and locations are extracted and labeled accordingly. Entity extraction goes one step further to identity relations and label each word in these sentences.

3.1.7 Intent classification :

Every domain specific assistant has a predefined set of abilities that can perform, depending the request or the user. Intent recognition refers to the task of understanding the user's request which should obviously be already known by the chatbot. According to (Taylor, 2004), to understand the meaning of the sentence, it is necessary to recognize all possible meanings of the utterance to choose the most appropriate one for the situation. In other words, upon receiving a new message, the chatbot has to be able to identify the target the user is trying to accomplish. This is modelled as a multi-classification problem whose classes are the set of the possible user intents.
Recurrent Neural Network (RNN) were built to solve this kind of problem, but since they face the vanishing gradient problem (Hochreiter, 1998), we used Long Short-Term Memory (LSTM) networks proposed in (Hochreiter & Schmidhuber, 1997), which have been widely used in this area.

3.2 Storage Unit (SUn) :

SUn is a dataset that all the important intents and entities are stored in for the sake of future machine learning algorithms

3.2.1 Response Manager Unit (RMU)

While talking about communication, a conversational agent must be able to reply on questions, a user provides.
Questions and responses could be either :

— **Open Domain :** the conversation has no limits, it can go into all kinds of directions. Until today, it represents a challenge for chatbots because it requires a huge knowledge of an infinite number of topics (Kamphaug *et al.*, 2018).

— **Closed Domain :** on the other hand, is a limited model because it is trying to achieve a specific goal. Meanwhile, users can talk about anything they want, but the model is not supposed to handle all the cases and normally, users should be aware about that (Kamphaug *et al.*, 2018).

And there are two types of response generation models :

— **Retrieval-based model :** use a list of predefined responses from which to pick the most appropriate answer. Retrieval systems are not able to generate new text to provide answers. One important point is that, through this model, chatbots can control every answer and avoid inappropriate replies (Kamphaug *et al.*, 2018).

— **Generative-based model :** generate new replies without the need for predefined responses. Responses are produced easily through a well trained model (Kamphaug *et al.*, 2018).

→ **Our solution is a rules-based conversational agent.**

After finishing all NLP steps, we have to train our agent on delivering responses. The training is applied on a customised data generated from scratch.

To classify responses, we used decision trees algorithm (Quinlan, 1993). To solve the problem of anxiety emotion detection we trained our agent on a set of scenarios composed of several questions and answers that help to supervise the user's status. If an anxiety emotion is detected through the responses given by the user, our conversational agent will take this situation into consideration ,in its future responses and may send reassurance messages.

Users may ask about numerous information concerning COVID-19 (Sohrabi *et al.*, 2020) that we illustrated them in Figure 5

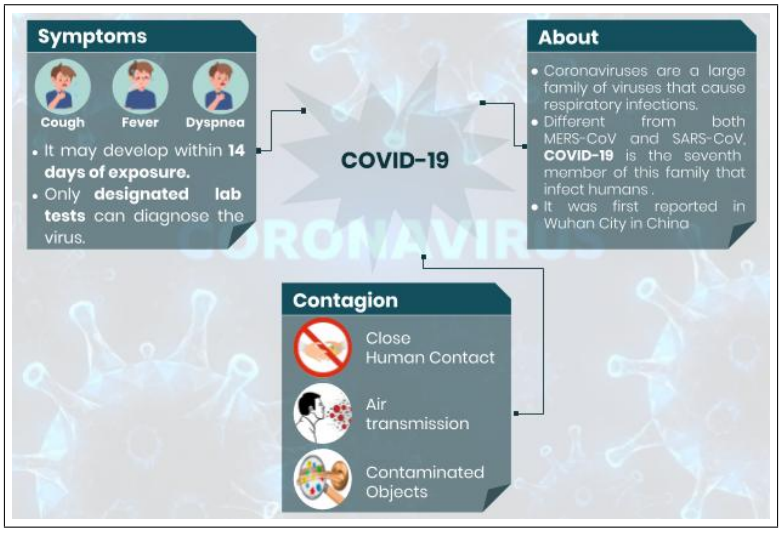


FIGURE *5 – Information about COVID-19*

4 Implementation

The development of this project was done on a Computer with Intel(R) Core(TM) i5-8250U Central Processing Unit (CPU) 3.40 GHz and 8.00-GB RAM. We have tested our chatbot android application on a Smart Phone with Hisilicon Kirin 710F CPU, 4.00-GB RAM and 2340x1080 Resolution.

The main software tools used are listed below :

— 64-bit Kubuntu 18.04 [4]

— Android Studio 6.0 [5]

— Visual Studio Code Version 1.32.3 [6]

The language chosen for our solution back-end was Python [7] version 3.6.7 for its compatibility with micro frameworks such as Flask [8] and modern deep learning libraries such as

4. https://kubuntu.org/
5. https://developer.android.com/studio/
6. https://code.visualstudio.com/
7. https://www.python.org/downloads/release/python-367/
8. http://flask.pocoo.org/

Keras [9]. We also used the open source conversational artificial intelligence (AI) framework Rasa [10].

Our proposal is a cloud based mobile application used by ordinary citizens. Each unit of our proposed architecture is introduced as web services.

Our solution is still under development, however Figure 6 shows a primer prototype of our mobile conversational agent.

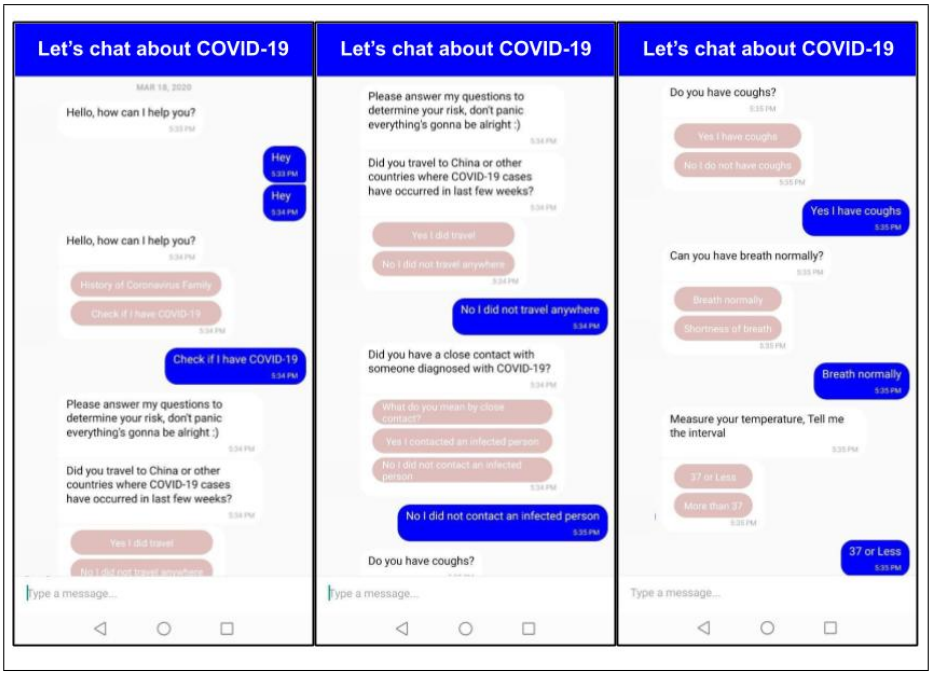


FIGURE *6 – The mobile conversational agent's prototype*

5 Conclusions and Future Works

In this paper we tried to introduce our proposed approach entitled mobile conversational agent for COVID-19 post quarantine psychological assistance. So presented the different units that compose our conversational agent. While NLP task is solved through several steps done successively such as tokenization, lower casing, PoS Tagging, stop words filtering, word embedding, etc, we managed the anxiety detection problem through a well determined set of questions delivered by the Response Manager Unit. Our solution is still under development, we aim at developing a deep leaning model to handle the problem of anxiety emotion detection.

Références

ALIMADADI A., ARYAL S., MANANDHAR I., MUNROE P., JOE B. & CHENG X. (2020). Artificial intelligence and machine learning to fight covid-19. volume 52.

BARSTUĞAN M., ÖZKAYA U. & ÖZTÜRK (2020). Coronavirus (covid-19) classification using ct images by machine learning methods.

BENGIO Y., DUCHARME R. & VINCENT P. (2000). A neural probabilistic language model. volume 3, p. 932–938.

9. https://keras.io/
10. https://rasa.com/

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2016). Enriching word vectors with subword information. volume abs/1607.04606.

DEVLIN J., CHANG M., LEE K. & TOUTANOVA K. (2018). BERT : pre-training of deep bidirectional transformers for language understanding. volume abs/1810.04805.

GOYAL K., CHAUHAN P., CHHIKARA K., GUPTA P. & SINGH M. P. (2020). Fear of covid 2019 : First suicidal case in india ! *Asian Journal of Psychiatry*, **49**, 101989.

GOZES O., FRID-ADAR M., GREENSPAN H., BROWNING P., ZHANG H., JI W., BERNHEIM A. & SIEGEL E. (2020). Rapid ai development cycle for the coronavirus (covid-19) pandemic : Initial results for automated detection patient monitoring using deep learning ct image analysis.

HOCHREITER S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. volume 6, p. 107–116.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long short-term memory. volume 9, p. 1735–1780.

HOSSAIN M. S. & MUHAMMAD G. (2017). An emotion recognition system for mobile applications. volume 5, p. 2281–2287.

KAMPHAUG Å., GRANMO O.-C., GOODWIN M. & ZADOROZHNY V. I. (2018). Towards open domain chatbots—a gru architecture for data driven conversations. In S. DIPLARIS, A. SATSIOU, A. FØLSTAD, M. VAFOPOULOS & T. VILARINHO, Eds., *Internet Science*, p. 213–222, Cham : Springer International Publishing.

KRIPKE S. (1980). : Harvard University Press.

KUMAR D. & JOSAN G. S. (2010). Part of speech taggers for morphologically rich indian languages : A survey.

MIIKKULAINEN R. & DYER M. G. (1991). Natural language processing with modular pdp networks and distributed lexicon. volume 15, p. 343 – 399.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. volume abs/1301.3781.

O'SHEA K. (2012). An approach to conversational agent design using semantic sentence similarity. volume 37, p. 558–568.

O'SHEA K. (2014). Natural language scripting within conversational agent design. volume 40, p. 189–197.

PENNINGTON J., SOCHER R. & MANNING C. (2014). Glove : Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1532–1543 : Association for Computational Linguistics.

QUINLAN J. R. (1993). C4.5 : Programs for machine learning. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

RAO A. S. S. & VAZQUEZ J. A. (2020). Identification of covid-19 can be quicker through artificial intelligence framework using a mobile phone-based survey in the populations when cities/towns are under quarantine. p. 1–18 : Cambridge University Press.

SILVA B. M., RODRIGUES J. J., DE LA TORRE DÍEZ I., LÓPEZ-CORONADO M. & SALEEM K. (2015). Mobile-health : A review of current state in 2015. volume 56, p. 265 – 272.

SOHRABI C., ALSAFI Z., O'NEILL N., KHAN M., KERWAN A., AL-JABIR A., IOSIFIDIS C. & AGHA R. (2020). World health organization declares global emergency : A review of the 2019 novel coronavirus (covid-19). volume 76, p. 71 – 76.

TAYLOR J. (2004). Toward computational recognition of humorous intent.

WANG L. & WONG A. (2020). Covid-net : A tailored deep convolutional neural network design for detection of covid-19 cases from chest radiography images.

WEBSTER J. J. & KIT C. (1992). Tokenization as the initial phase in nlp.

Diviser pour mieux classifier

Yves Mercadier[1], Jérôme Azé[1], and Sandra Bringay[1,2]

[1] LIRMM, Univ Montpellier, CNRS, Montpellier, France
name@lirmm.fr, http://www.lirmm.fr/

[2] Université Paul-Valéry Montpellier 3, Montpellier, France

Résumé : L'information médicale est présente dans différentes sources textuelles comme les dossiers médicaux informatisés, la littérature biomédicale, les médias sociaux, etc. Exploiter l'ensemble de ces sources pour en extraire de l'information utile représente un véritable défi. Dans ce contexte, la classification de textes est une tâche importante. Récemment, les classifieurs profonds ont montré leur capacité à obtenir de très bons résultats sur des tâches de classification mono-étiquette [1] mais leurs résultats dépendent généralement de la quantité de données utilisée pour l'entraînement. Dans cet article, nous proposons une nouvelle approche pour augmenter les données textuelles. Nous avons testé cette approche sur 5 jeux de données réels en la comparant avec les principales approches de la littérature. Notre proposition améliore les performances dans les configurations que nous avons pu mettre en œuvre.

Natural language processing, Document classification, Textuel data augmentation.

Abstract : Medical information can be found in a variety of textual sources such as computerized medical records, bio-medical literature, social media, etc. Exploiting all of these sources to extract useful information is a real challenge. In this context, the classification of texts is an important task. Recently, deep classifiers have shown their ability to perform very well on single label classification tasks but their results are generally dependent on the amount of data used for training. In this article, we propose a new approach to augment textual data. We tested this approach on 5 real data sets by comparing it with the main approaches in the literature. Our proposal improve the results in every configuration we have been able to implement.

1 Introduction

L'information médicale est présente dans différentes sources textuelles comme les dossiers médicaux informatisés, la littérature biomédicale, les médias sociaux, etc. Récemment, les classifieurs profonds ont montré leur capacité à obtenir de très bons résultats pour cette tâche. Cependant, ces résultats dépendent généralement de la quantité de données utilisée pour l'entraînement.

Dans cet article, nous nous intéressons à l'augmentation de données qui peut s'avérer efficace sur de petits jeux de données. L'augmentation de données exploite des données en quantité limitée et transforme les échantillons existants pour en créer de nouveaux. Plus précisément, il s'agit d'injecter de la connaissance en prenant en compte les propriétés invariantes des données par rapport à certaines transformations. Les données augmentées peuvent ainsi couvrir un espace d'entrée inexploré, éviter le sur-apprentissage et améliorer la généralisation du modèle.

Cette technique s'est avérée efficace pour des tâches de classification d'images notamment lorsque la base de données d'entraînement est limitée. Par exemple, (He *et al.*, 2015) ont montré que les changements mineurs dus à l'échelle, au recadrage, à la déformation, à la rotation, etc. ne modifient pas les étiquettes car ces changements sont susceptibles de se produire dans des observations du monde réel. Cependant, les transformations qui préservent les étiquettes pour les données textuelles ne sont pas aussi évidentes, ni aussi intuitives.

Dans cet article, nous présentons une technique d'augmentation de données appliquées aux textes que nous appellerons DAIA (**d**ata **a**ugmentation and **i**nference **a**ugmentation). Nous

1. Un exemple appartient à une seule classe et une seule étiquette.

évaluerons DAIA pour 6 jeux de données de santé de nature différente en classification mono-étiquette et nous montrerons une amélioration par rapport à l'état de l'art, particulièrement sur les petites bases de données.

2 État de l'art

L'augmentation de données a été utilisée avec succès, dans le domaine de l'analyse d'images. Par exemple, Perez & Wang (2017) ont comparé plusieurs techniques simples, telles que le recadrage, la rotation et le retournement d'images avec des techniques plus évoluées comme les GAN pour générer des images de différents styles ou encore des approches d'augmentation par réseau neuronal apprenant les augmentations qui améliorent le plus un classifieur. On distingue quatre approches principales que nous allons décrire ci-dessous :

Des approches utilisant des ressources sémantiques ont tout d'abord été proposées. Par exemple, Zhang & LeCun (2015) ont utilisé un thésaurus pour remplacer les mots par leurs synonymes afin de créer un jeu de données augmenté pour une tâche de classification de textes. Cette augmentation a même diminué les performances dans certains cas.

Des approches inspirées des distorsions que l'on peut rajouter dans les images, ont été également appliquées aux textes. Pour une tâche de classification, Wei & Zou (2019), dans la méthode EDA, augmentent le nombre d'échantillons en supprimant, permutant un mot ou en le remplaçant par un synonyme. Les expérimentations sont faites avec des classifieurs de type CNN et RNN et montrent de réelles améliorations sur les petits jeux de données. Par exemple, l'exactitude progresse de 3% en moyenne avec le classifieur CNN pour 500 textes étudiés. Des approches se sont focalisées sur le choix des mots à modifier. Dans la méthode UDA, Xie *et al.* (2019) remplacent les mots à faible contenu informationnel, repérés avec un TF.IDF faible, par leur synonyme tout en gardant ceux qui ont des valeurs de TF.IDF élevées représentant les mots clés. Cette heuristique a été testée sur six corpus de textes, ce qui a permis de ramener l'erreur de classification dans l'intervalle 0.3% à 3%.

Des approches génératives de type GAN ont également été étudiées. Dans le domaine spécifique de la réponse visuelle aux questions (Visual Question Answering VQA), qui consiste à partir d'une image et d'une question sur cette image, à prédire la réponse à la question, Kafle *et al.* (2017) ont mis en place deux approches pour générer de nouvelles questions. Pour cela, ils utilisent des annotations sémantiques existantes combinées avec une approche générative utilisant des réseaux de neurones récurrents. Leurs deux méthodes améliorent de 1 à 2% l'exactitude en augmentant la variété et le nombre de questions.

Une dernière approche se base sur l'augmentation de données textuelles à l'aide de rétro-traduction (back-translation) (Sennrich *et al.*, 2015). Il s'agit de traduire un exemple dans une langue donnée puis de retraduire la traduction obtenue dans la langue initiale. Yu *et al.* (2018) expliquent ainsi que la traduction inversée génère des paraphrases en préservant la sémantique des phrases d'origine. Sur le jeu de données SQAD (Stanford Question Answering Dataset), ils obtiennent une amélioration de 3% de l'exactitude. Cependant Shleifer (2019) a montré que la technique de traduction inversée n'apportaient que peu d'améliorations avec les classifieurs modernes comme UMLfit.

Dans cet article, nous allons nous focaliser ici sur une approche de distorsion, simple à mettre à mettre en place, ne nécessitant pas de ressource comme dans les approches sémantiques, ni de grandes quantités de calculs comme les approches génératives ou encore l'accès à des ressources externes comme les approches de traduction inversée. Une limite des approches de distorsion que l'on retrouve dans la littérature est qu'elles ne préservent pas l'ordre des mots dans les phrases. Les exemples alors générés s'éloignent de ceux que l'on pourrait retrouver dans le monde réel. Dans l'approche DAIA, nous allons décrire une approche pour découper les phrases du jeu d'apprentissage en plusieurs séquences utilisées pour l'augmentation dans la phase d'entraînement. La même approche sera utilisée durant la phase de test sur les exemples à classifier en appliquant une technique de vote pondéré.

Nous montrerons dans les expérimentations que cette approche améliore les résultats d'une classification de textes quel que soit le type de classifieurs profond utilisé comme Bert (Devlin *et al.*, 2019), RoBERTa (Liu *et al.*, 2019), Albert (Lan *et al.*, 2019), DistilBert (Sanh

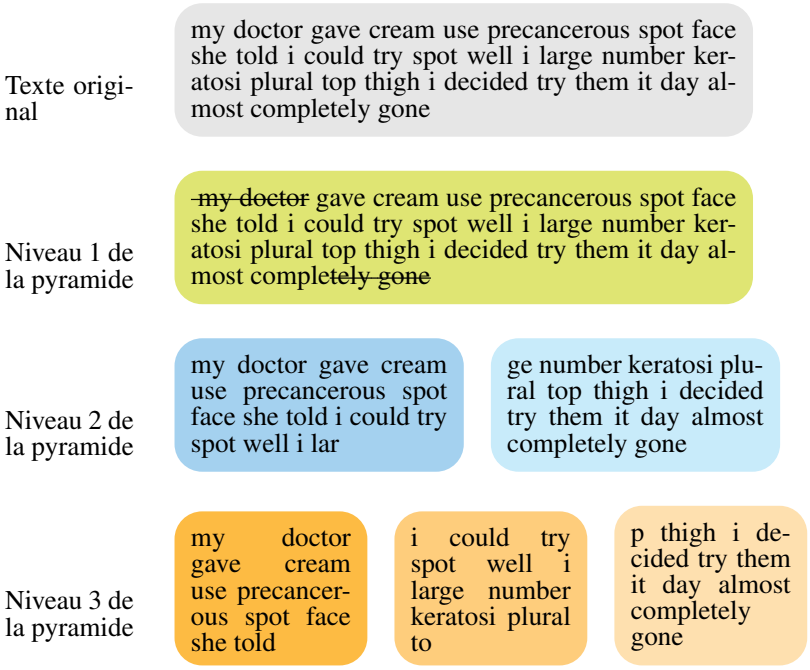


Figure 1 – Description de la découpe pyramidale. La figure illustre un découpage en 3 niveaux, permettant ainsi d'obtenir 6 nouveaux documents (plus le texte original).

et al., 2019) ou encore SciBert (Beltagy *et al.*, 2019). Nous comparerons également les expérimentations sur notre proposition DAIA avec les approches par distorsion UDA (Xie *et al.*, 2019) et EDA (Wei & Zou, 2019), l'approche générative TextGen (Gupta, 2019) ou encore la traduction inversée (Shleifer, 2019).

3 DAIA

Notre proposition DAIA s'articule en deux parties : sur la phase d'entraînement (DA : Data Augmentation) et sur la phase de test (IA : Inference Augmentation).

Augmentation des données : Dans la phase d'entraînement, nous augmentons la quantité de donnée en découpant le texte initial de chaque échantillon. Nous cherchons à produire de nouveaux échantillons sans modifier la relation d'ordre existant dans la succession des mots de chaque phrase afin de ne pas diminuer la qualité d'apprentissage de description des plongements de mots. Nous proposons une découpe pyramidale. Cette nouvelle découpe se décline en n niveaux. L'augmentation pour le niveau 1 se fait par la découpe symétrique des bordures du texte. L'augmentation pour le niveau 2 se fait en découpant le texte en deux parties égales et en rajoutant l'augmentation obtenue au niveau 1. L'augmentation pour le niveau i se fait en découpant le texte en i parties égales et en rajoutant l'augmentation de niveau $i-1$ comme illustré par la figure 1. Pour chaque texte, nous obtenons ainsi $\frac{n\times(n+1)}{2}$ nouveaux documents étiquetés.

Inférence : La phase de test consiste à prédire des étiquettes pour de nouveaux textes à partir du modèle appris pendant la phase d'apprentissage. Nous découpons le texte de l'exemple à classer en suivant le même protocole que celui de la phase d'entraînement, puis nous donnons au classifieur l'ensemble des séquences de textes générées. Pour chaque séquence, le classifieur renvoie un ensemble de prédictions par découpe qui sont agrégées par vote pondéré (somme des valeurs pour chaque classe prédites de chaque élément de l'ensemble). Ainsi, pour chaque texte à classer, nous obtenons une valeur par classe et retenons l'étiquette associée à la valeur maximale des prédictions obtenues par classe.

4 Présentation des expérimentations

4.1 Jeux de données

Afin de montrer la généricité de notre approche, nous avons choisi cinq jeux de données dans le domaine de la santé décrits dans le tableau 1. Les deux premiers correspondent à des publications médicales. Les trois autres jeux correspondent à des textes écrits par des patients. Il est important de noter que ces jeux de données ne sont pas équilibrés en classe.

Jeu de données	# classes	# documents	Nombre moyen de mots par texte	Tâche de classification
PubMed 200k RCT	5	2 211 861	26.22	analyse de texte scientifique
WHO COVID-19	4	26 909	166	analyse de provenance
Drugs.com	10	53 766	85.58	analyse de sentiment
eR anorexie	2	84 834	38.24	analyse de sentiment
eR dépression	2	531 394	36.76	analyse de sentiment

Table 1 – Jeux de données utilisés dans nos expériences pour la classification de textes.

PubMed 200k RCT [2] (Dernoncourt & Lee, 2017) est une base contenant environ 200 000 résumés d'articles portant sur des essais contrôlés randomisés, totalisant plus de 2 millions de phrases. Chaque phrase est étiquetée selon sa signification dans le résumé (contexte, objectif, méthode, résultat ou conclusion).

WHO COVID-19 [3], en réponse à la pandémie de COVID-19, l'institut *Allen for AI* s'est associé à des groupes de recherche de premier plan pour distribuer un jeu de données composé de plus de 29 000 articles, dont plus de 13 000 textes dans leur intégralité, sur le COVID-19 et la famille des coronavirus. Nous avons utilisé dans cette étude uniquement les articles ayant un résumé. Chaque texte est étiqueté selon sa source : CZI, PMC, biorxiv, medrxiv.

Drugs.com [4] (Gräßer *et al.*, 2018) correspond à des avis de patients sur des médicaments, associés à des maladies. Les données ont été obtenues en analysant les sites pharmaceutiques en ligne. Chaque texte est étiqueté selon une note de 1 à 10 correspondant à la satisfaction des patients.

Les jeux de données eR Depression et eR Anorexie sont issus du challenge CLEF eRisk 2018 [5]. Les textes correspondent à des messages d'utilisateurs dans des réseaux sociaux étiquetés selon les classes dépression/non dépression et anorexie/non anorexie (Ragheb *et al.*, 2018).

4.2 Pré-traitements des données

Pour chaque jeu de données, nous avons appliqué les pré-traitements suivants : suppression des ponctuations, des caractères spéciaux, des mots vides, changement des majuscules en minuscules et lemmatisation [6]. Chaque texte est associé à une seule étiquette. Ces étiquettes sont utilisées comme sortie de la classification.

4.3 Comparaison à d'autres approches de l'état de l'art

Nous comparons notre proposition à trois méthodes de l'état de l'art que nous décrivons en détail ci-dessous.

Pour les approches par distorsion sémantique de mot, nous avons considéré les approches EDA et UDA.

2. https://github.com/Franck-Dernoncourt/pubmed-rct
3. https://pages.semanticscholar.org/coronavirus-research
4. https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29
5. https://early.irlab.org/2018/index.html
6. La lemmatisation est faite par le module python Nltk avec l'aide du dictionnaire WordNet

Pour EDA, nous mettons en œuvre les quatre techniques simples d'augmentation de données décrites par (Wei & Zou, 2019) à savoir le remplacement par des synonymes, l'insertion aléatoire, l'échange aléatoire, la suppression aléatoire. Pour cela, nous avons utilisé le dictionnaire de synonymes Wordnet de NLTK [7].

Pour UDA, nous procédons comme Xie *et al.* (2019) qui repèrent ces mots comme étant ceux négativement corrélés à leur score TF.IDF. Pour cela, ils définissent la probabilité $min(p(C-TF.IDF(x_i))/Z, 1)$, où p est un hyper-paramètre contrôlant la variation d'augmentation, C est le score maximum de TF.IDF pour les mots x_i d'un texte x et $Z = \sum_i (C-TF.IDF(x_i))/|x|$. Pour nos expérimentations, nous avons choisi $p = 0.9$. Xie *et al.* (2019) ne modifient pas les mots clés, repérés par leur fréquence. Les mots repérés qui ne sont pas des mots clés sont remplacés par un des mots non essentiels du corpus.

Comme les générateurs de textes de l'état de l'art utilisés pour l'augmentation des données ont été dépassés en qualité sémantique par le générateur GPT2 (Gupta, 2019), nous avons utilisé ce dernier selon le protocole suivant. Pour chaque texte, un texte augmenté est construit par concaténation de deux parties. Une partie initiale, la graine, correspondant à la moitié du texte original et une deuxième partie obtenue en utilisant le générateur GPT2 ayant pris en entrée la graine.

L'augmentation de données par traduction inversée (Shleifer, 2019) consiste à produire des paraphrases conservant globalement la sémantique de la phrase initiale. Nous utilisons le service web de translation Yandex [8] afin de produire ces paraphrases. Nous procédons à une première traduction des textes en japonais puis nous retraduisons ces textes en anglais et enfin nous les étiquetons avec le label original du texte.

4.4 Protocole d'évaluation, Hyper-paramétrage et entraînement

Nous utilisons le module python Mantéïa pour l'implémentation [9]. Nous procédons comme suit pour chaque jeu de données mono-étiquette décrit dans la section 4.1. Nous extrayons 5 000 textes aléatoirement en respectant la pondération des classes. Tous les résultats de cette étude ont été calculés sur la moyenne d'une 4-validation croisée stratifiée.

L'ensemble des hyper-paramètres des réseaux est fixé pour l'ensemble des jeux de données. Le taux d'apprentissage est fixé à 0.00001. Les optimiseurs de calcul du gradient sont de type *Adam weight*. De plus, nous utilisons une mise à jour du taux d'apprentissage linéaire avec une amélioration du départ. Les expérimentations sont faites sur deux GPU de type GeForce RTX.

5 Résultats des expérimentations

5.1 Impact du classifieur et comparaison à l'état de l'art

Nous travaillons dans la suite sur le jeu de données drugs.com avec un échantillon de 5 000 textes. La baseline correspond à l'utilisation d'un classifieur seul sans augmentation de données. Les classifieurs comparés sont les classifieurs les plus performants de la littérature : Bert (Devlin *et al.*, 2019), RoBERTa (Liu *et al.*, 2019), Albert (Lan *et al.*, 2019), DistilBert (Sanh *et al.*, 2019) ou encore SciBert (Beltagy *et al.*, 2019). La méthode DAIA est appliquée ici pour l'augmentation des données pendant la phase d'apprentissage mais également pendant la phase de test comme détaillé dans la section 3. La découpe pyramidale est supérieure aux autres approches de la littérature comme EDA et à UDA ou encore la traduction inversée, pour les réseaux Roberta et Xlnet. Combinée avec l'inférence augmentation, DAIA sur-performe pour tous les classifieurs. Les réseaux les plus importants en termes de nombre

7. https://www.nltk.org/howto/wordnet.html

8. https://translate.yandex.com/

9. `https://github.com/ym001/Manteia/blob/master/notebook/notebook_Manteia_classification_augmentation_run_in_colab.ipynb`

Table 2 – Étude en fonction du classifieur et comparaison aux approches de la littérature

Classifieur	Roberta	Bert	Xlnet	Albert	Distilbert	Scibert
Baseline sans DA	0.3661	0.3637	0.3760	0.3200	0.3601	0.3392
Distorsion EDA	0.38	0.3669	0.3712	0.3226	0.3689	0.3510
Distorsion UDA	0.3811	0.3632	0.3525	0.3084	0.3660	0.3327
Génération de textes GPT2	0.3384	0.3252	0.3425	0.3122	0.3204	0.3078
Traduction inversée	0.3798	0.3462	0.3728	0.3426	0.3584	0.3361
DA - découpe pyramidale	0.3877	0.3660	0.3762	0.334	0.3688	0.3590
DAIA	**0.3931**	**0.3755**	**0.3786**	**0.3444**	**0.3693**	**0.3625**

de neurones sur-performent les autres et parmi eux, c'est le réseau Roberta qui dépasse les autres en terme d'exactitude.

5.2 Impact du nombre de classes

Dans le tableau 3, nous étudions les performances de DAIA en fonction du nombre de catégories pour les cinq corpus de l'étude. Nous avons retenu Roberta comme classifieur avec un échantillonnage de 500 textes. On remarque une amélioration apportée par DAIA plus importante pour les jeux ayant un nombre de classes élevé comme Drugs.com (10).

Table 3 – Impact du nombre de classes

classifieur	Drugs.com	PubMed 200k RCT	WHO COVID-19	eR Depression	eR anoxerie
Baseline (Roberta)	0.2846	0.6413	0.9319	0.8926	0.9079
DAIA	**0.316** (+11.03%)	**0.6513** (+1.56%)	**0.938** (+0.65%)	**0.9066** (+1.57%)	**0.9153** (+0.82%)

5.3 Impact de la quantité de données dans le jeu d'apprentissage

Nous travaillons ici avec comme baseline le classifieur Roberta sans augmentation de données. Nous montrons dans la figure 2 que DAIA apporte une amélioration pour toutes les tailles des jeux de données étiquetées Drugs.com (à gauche) et eR anorexie (à droite). En particulier, dès les jeux de données de 500 textes, on observe une amélioration de 2.7%. L'impact se réduit quand les jeux atteignent la taille des 5 000 textes.

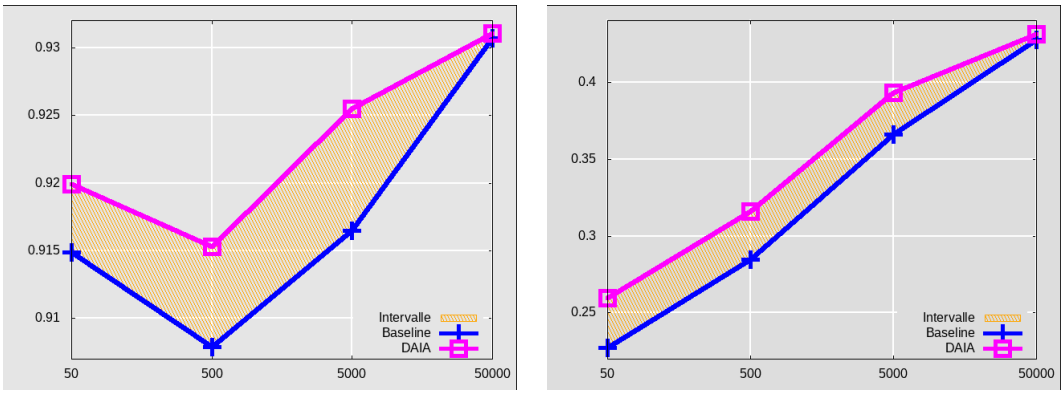


Figure 2 – Étude en fonction de la taille du corpus.

5.4 Impact de la taille de séquence retenue

Dans la figure 3, nous étudions les variations de DAIA selon la taille des découpes obtenues et la longueur des textes en entrée pour le jeu de données drugs.com. mais des résultats équivalents ont été obtenus sur les autres jeux de données de l'étude. Un maximum semble atteint

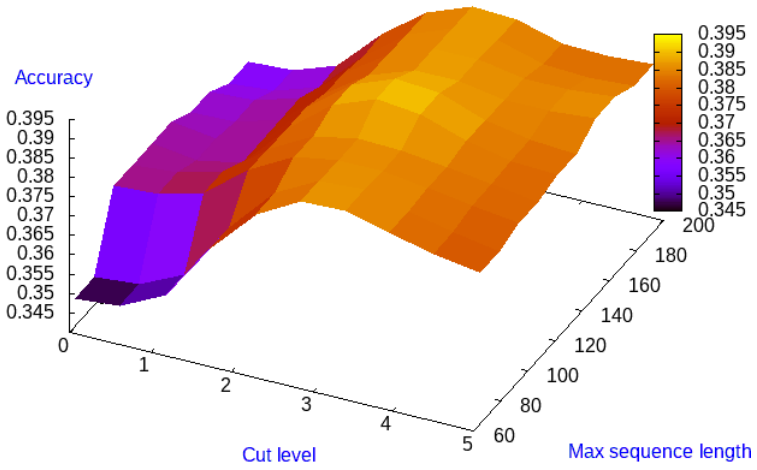


Figure 3 – Étude en fonction du niveau de découpe pyramidal et de la longueur maximale de séquence en entrée du réseau.

pour un niveau de découpe trois associé à une longueur de séquence de 128 sous-mots appartenant au vocabulaire du réseau de neurones. La taille du texte en entrée a finalement peu d'impact dans l'intervalle étudié avec un très léger maximum pour la valeur de 128 sous-mots.

6 Conclusion

Dans cette étude, nous avons proposé une nouvelle méthode pour augmenter les données textuelles, qui consiste à découper les textes en plusieurs segments pour augmenter la variété des textes d'entraînement tout en préservant la qualité de l'apprentissage des plongements de mots. Notre approche DAIA a amélioré les performances pour cinq jeux de données dans le domaine de la santé, pour six différents classifieurs et a été comparée avec succès avec les principales approches de la littérature. L'impact du choix de la langue sur la retro-traduction doit être évalué. Une expérimentation supplémentaire associée à notre approche consisterait à faire varier les découpes sur les textes, selon que l'on s'efforce de coller au plus près des textes initiaux ou au contraire que l'on se permette de générer des textes peu plausibles mais qui permettraient néanmoins d'améliorer le classifieur. Par ailleurs, il est possible de conserver le sens sans conserver l'ordre de mots. On pourrait alors imaginer de nouvelles expérimentations consistant à conserver certains niveaux de l'articulation syntaxique plutôt que l'ordre des mots afin de générer plus de variabilité tout en conservant la sémantique de la phrase. Nous aimerions également poursuivre ces recherches vers d'autres types de données, comme les images ou les sons et nous intéresser à des tâches plus complexes comme la classification multi-étiquettes. Enfin, il serait également très intéressant d'appliquer DAIA aux heuristiques de deep active learning dans le domaine médical.

References

BELTAGY I., LO K. & COHAN A. (2019). Scibert: A pretrained language model for scientific text. In K. INUI, J. JIANG, V. NG & X. WAN, Eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, p. 3613–3618: Association for Computational Linguistics.

DERNONCOURT F. & LEE J. Y. (2017). Pubmed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In G. KONDRAK & T. WATANABE, Eds., *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017, Volume 2: Short Papers*, p. 308–313: Asian Federation of Natural Language Processing.

DEVLIN J., CHANG M., LEE K. & TOUTANOVA K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In J. BURSTEIN, C. DORAN & T. SOLORIO, Eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, p. 4171–4186: Association for Computational Linguistics.

GRÄSSER F., KALLUMADI S., MALBERG H. & ZAUNSEDER S. (2018). Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. In P. KOSTKOVA, F. GRASSO, C. CASTILLO, Y. MEJOVA, A. BOSMAN & M. EDELSTEIN, Eds., *Proceedings of the 2018 International Conference on Digital Health, DH 2018, Lyon, France, April 23-26, 2018*, p. 121–125: ACM.

GUPTA R. (2019). Data augmentation for low resource sentiment analysis using generative adversarial networks. *CoRR*, **abs/1902.06818**.

HE K., ZHANG X., REN S. & SUN J. (2015). Deep residual learning for image recognition. *CoRR*, **abs/1512.03385**.

KAFLE K., YOUSEFHUSSIEN M. A. & KANAN C. (2017). Data augmentation for visual question answering. In J. M. ALONSO, A. BUGARÍN & E. REITER, Eds., *Proceedings of the 10th International Conference on Natural Language Generation, INLG 2017, Santiago de Compostela, Spain, September 4-7, 2017*, p. 198–202: Association for Computational Linguistics.

LAN Z.-Z., CHEN M., GOODMAN S., GIMPEL K., SHARMA P. & SORICUT R. (2019). Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, **abs/1909.11942**.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, **abs/1907.11692**.

PEREZ L. & WANG J. (2017). The effectiveness of data augmentation in image classification using deep learning. *CoRR*, **abs/1712.04621**.

RAGHEB W., MOULAHI B., AZÉ J., BRINGAY S. & SERVAJEAN M. (2018). Temporal mood variation: at the CLEF erisk-2018 tasks for early risk detection on the internet. In L. CAPPELLATO, N. FERRO, J. NIE & L. SOULIER, Eds., *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018*, volume 2125 of *CEUR Workshop Proceedings*: CEUR-WS.org.

SANH V., DEBUT L., CHAUMOND J. & WOLF T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, **abs/1910.01108**.

SENNRICH R., HADDOW B. & BIRCH A. (2015). Improving neural machine translation models with monolingual data. *CoRR*, **abs/1511.06709**.

SHLEIFER S. (2019). Low resource text classification with ulmfit and backtranslation. *CoRR*, **abs/1903.09244**.

WEI J. W. & ZOU K. (2019). EDA: easy data augmentation techniques for boosting performance on text classification tasks. *CoRR*, **abs/1901.11196**.

XIE Q., DAI Z., HOVY E. H., LUONG M. & LE Q. V. (2019). Unsupervised data augmentation. *CoRR*, **abs/1904.12848**.

YU A. W., DOHAN D., LUONG M., ZHAO R., CHEN K., NOROUZI M. & LE Q. V. (2018). Qanet: Combining local convolution with global self-attention for reading comprehension. *CoRR*, **abs/1804.09541**.

ZHANG X. & LECUN Y. (2015). Text understanding from scratch. *CoRR*, **abs/1502.01710**.

Predictive Patient Care: Visualize and Interpret Models Decisions Application to Medication Adherence

Thomas Janssoone[1], Clémence Bic[1], Nathan Casals[1], Pierre Rinder[1], Pierre Hornus[1] and Dorra Kanoun[2]

[1] SÊMEIA COMPANY, 9 cour des petites écuries, 75011 Paris, France
Corresponding author: tjanssoone@semeia.io

[2] CLINIQUE PASTEUR , Toulouse, France

Abstract : More and more machine learning models are applied to solve health-care issues. From cancer detection in medical images to text mining of medical reports through risk prediction from Electronic Health Records (EHR), the abundance of data is opening a new era to health-care. Yet, many challenges remain such as the rare and weak labeling of the data, their privacy or the ability to explain and justify the decision of a model. This last part is the focus of this paper which illustrates it with our work applied to the prediction of bad medication persistence that has been discussed with gynecologists and oncologists. Persistence in medicine is a measure of how well a patient follows their treatment. The World Health Organization reports[1] that not following the medication plan is actually a major issue as it severely compromises the efficiency of long-term therapy and increases the cost of health services. Indeed, in developed countries, only about 50% of patients with chronic diseases correctly follow their treatments.

This paper reports our work in explaining model decisions predicting patient drug consumption in breast cancer treatments from medical claims. We use reimbursement records from the French Health System to analyze patient care paths. These raw data are processed to be analyzed by different machine learning methods to evaluate the risk of an illegitimate drop out of the treatment. While obtaining good results with these predictive models, we now have to convince the medical staff of their righteousness by explaining how they made their decision. We detail and explain the different metrics and tools we use before discussing their limitations through the feedback of doctors who tested them.

Mots-clés : Persistence, Adherence, Machine Learning Interpretability, EHR SNDS

1 Introduction

With more and more data available, machine learning approaches are becoming prevalent to handle healthcare issue(9). These methods their efficiency in tasks such as breast cancer detection in images (8) or prediction of adverse events for EHR analysis (4). Yet, these machine learning-based solutions can be misused(2) or even lead to a decrease of performance as it can disturb medical experts with too many alarms (18). This observation underlines the need for explanations and visualizations of models decision to convince the medical staff and avoid errors thanks to their feedback. We follow up on our work detailed in Janssoone *et al.* (10) showing our ability to predict non-compliance in oral cancer therapy administered to the patient. As they are widely used for long-term illness such as cancer, adherence to medication is becoming a major health issue as *"non-adherence (or non-persistence) can lead to increased morbidity, mortality, and healthcare costs"*(11). The term adherence is a composite concept that generally encompasses the concept of compliance and persistence. Therapeutic compliance describes, in general, to what extent the patient complies with the prescription (diet, dosage, etc.). Persistence defines as the respect of the duration of the treatment until its term, and this, without interruption of this one.

[1] http://www.who.int/chp/knowledge/publications/Persistence_full_report.pdf

Our work focuses on non-persistence in medicine as the measure of the fact that a patient stops following a treatment before the end of the recommended period. We aim to detect an illegitimate drug drop-out during a treatment for a long term illness. Indeed, a common solution to prevent these illegitimate stops is to provide support to patients so knowing the risk of each patients at each time could improve the support. The major issue with these programs is that human interventions are effective but very expensive limiting their reach, while the use of digital technologies (notifications and explanations) is too generic and sometimes too intrusive (daily reminders) which leads to patients losing interest (15). Yet, today, French Health Services choose randomly the patients they call to check on them. To make sure these interventions target patient at risk of an illegitimate drug drop-out, we designed a predictive model based on the medical claims data of patients with breast cancer. We manage to evaluate for each of them a risk index reflecting their persistence to the treatment. This allows us to predict the most appropriate moments to notify the ones really needing help. To validate our results and to justify them to the medical staff, we use an algorithms to explain the decision made by the models. This allows to evaluate the effect of some characteristics of the patient care-path on the risk of non-adherence. We provide visualizations of these effects to medical staff that we will detail in this paper.

2 Related work

As non-observance is a major concern, many surveys have tried to identify their determining factors. Many studies use statistical tools to study this phenomenon. Hence, (6; 14; 5) look at several factors of non-persistence such as the increasing age of patients, the treatment complexity level (multiple drugs, injections, ...), the impact of patients' mental health (depressive episodes have a very negative impact) or the influence of the patient's entourage (support of their spouse, family, relatives, and the wider social environment). These studies provide a-priori indications for detecting risk profiles of non-Persistence. Using machine learning methods, Franklin *et al.* (7) underline the difficulty to use details about the patient (Census based especially) on top of clinically relevant characteristics to predict adherence. They evaluate different approaches, using logistic regression and boosted logistic regression, to define three categories of adherence predictors. Hence, they underline that using census information or transaction data leads to poor prediction but, they point out that using persistence observations during the first month significantly increases the accuracy of the results. Lo-Ciganic *et al.* (12) confirm this nuance on the weight of each adherence prediction variable using Medicaid database which contains claims data and they focused on diabete II medications ones. They use random survival forests highlights to find patient-specific persistence thresholds to discriminate between hospitalization risks. Here again, the major variables are linked to patient history and previous transactions in pharmacy. More recently, Shickel *et al.* (20) just released a method to used deep-learning methods on Electronic Health Records to asses illness severity. They obtained very good results to predict organ failure and propose an interpretability mechanism using self-attention.
Yet, to be used, such models need to justify their decision to convince the medical staff of the accuracy of the risk evaluation. This paper proposes a practical case to explore how visualizations can help to design an AI model to predict the risk of an illegitimate stop during a treatment and explain its decision process.

3 Methodology

3.1 Raw data: the SNDS database

The predictive models we developed are trained and tested on the reimbursement data of the unique French Health System (SNDS *Système national des données de santé, French administrative health care database*). SNDS is one of the largest structured databases of health data in the world. It contains reimbursement data of the French Health System, covering 99.8% of the French population ($\simeq$ 66 million persons). In our context, useful data are, for example,

hospitalizations, drug purchases or contextual patient information (age, government supports (CMU,ACS), geographic information, ...). The interest of this database is that it is very well structured and has no-bias in term of social background due to its universal coverage. More details can be found in Tuppin *et al.* (21) and its structure can be visualized here[2]. Previous work has already shown the value of massive data mining to aid diagnosis, either by taking all the information for a "static" approach (17), or, more recently, by also incorporating dynamic information (16).

Our study focuses on women's breast cancer on part of the SNDS data. The cohort of the study consists of women who meet the following criteria: (1) diagnosed with breast cancer, (2) having purchased at least one of the following molecules for the studied period: *Anastrozole*, *Capecitabine*, *Cyclophosphamide*, *Etoposide*, *Everolimus*, *Exemestane*, *Lapatinib*, *Letrozole*, *Megestrol*, *Melphalan*, *Tamoxifen*, *Toremifene* or *Vinorelbine*. We focused on these medications as they must be taken for a long time ($\simeq$ 10 years). Extraction (SNDS file n°736952, CNIL authorization decision DR 2019-322) concerns consumption between 2013 and 2015 and is made up of three main categories:

- Pharmacy transactions (molecule, number of doses, date, ...)
- Hospitalizations (diagnosis, start date, end date, ...)
- Patient information (age, department, date of the diagnostic of eventual long-term illness (referred as *ALD*) , pathologies, CMU-C[3], ACS[3], ...)

We then obtain a set of claims data linked to a patient care-path, how to predict the risk of an illegitimate drug drop-out and justify this risk to the medical staff. We now detail how we processed these data to identify the period of illegitimate stops by evaluating the risk that a patient does not go back to the pharmacy while she is still undertaking a treatment.

3.2 Preprocessing of the data

The raw data are administrative records for the reimbursements of patients during their carepath. This means that it does not contains prescriptions or biology results but tracks of transactions in pharmacies or bill of hospitalizations (with proof of medical act to evaluate the total cost). The first main challenge is to turn these administrative records into data reflecting the care-paths of the patients to model their behavior and extract information from them. Our first process is then to retrieve the patient care-path from these indirect observations. To do so, we first analyze the data to restore the chronological set of events with their timestamps. To characterize them, we measured the delay between them to find the normal chain of events. Working with feedbacks from medical doctors, we processed raw data into two forms. First, phases of treatment which can be seen as the big steps occurring during a patient care-path (in our case, chemotherapy, radiotherapy, and continuous intake of a drug). Secondly, transaction-centered which focus on every time the patient buy some in a pharmacy. A *phase* is a period of continuous intake of a molecule or hospitalizations for chemotherapy or radiotherapy. This allows the reconstruction of the patient's care path. For each phase, the following data is calculated:

- Start and end dates, number of intakes or hospitalizations, molecule or type of hospitalization
- End of treatment type (switch, death, switch to palliative care, right censorship or illegitimate stop)
- Patient information (comorbidities, number of consultations in the first year of treatment, age, health financial support (CMU/ACS), type of healthcare system (*regime*), number of consultations with doctors and nurses, time since the declaration of the long-term illness, department of residence)

[2] `https://drees.shinyapps.io/dico-snds/`

[3] CMU and ACS are French support for a person with low-income

- Interventions on the breast (mastectomy) during the three months before the studied phase

The *transaction centered* approach looks at each visit to the pharmacy as our previous analysis has shown (10), the drop out rate evolves over time and according to previous events in the patient's care path. For each purchase in a pharmacy, the input details 1- insight about the current transaction and the 9 previous one of drug used for treating breast cancer (Tamoxifen, Exemestane,...) with padding if needed, 2- hospitalizations occurring in between, 3- Patient information (comorbidities, age, ...). We then obtain two types of information: static ones with patient information (similar to phases), and dynamic ones which contains the five last transactions with details about the current medication, time since the previous transaction, hospitalizations, consultations with doctors and nurses. To label our data, we followed the recommendation oncologists: the criterion for identifying a drop-out is the existence of a two-months period plus hospitalizations after the median time covered by the last purchase without a refill of the molecule. This approach has been validated by doctors as the medication must be daily taken with a fixed-dose posology. The date of the last theoretical dose is obtained by calculating the median interval between two purchases of the molecule or two hospitalizations of the same type: this median behavior is considered to be in conformity with the dosage. The end of this period after the last box purchased corresponds to the date of the last theoretical take. Thus, the median time is 30 days between two box purchases of 30 doses of Tamoxifen. For medication, days of hospitalization are excluded from this period as the drug is then provided by the medical staff. The data can be labeled with one of the legitimate stops (death, switch of treatment, some kind of serious cardiac issue or beginning of palliative care) if this event occurs less than two months after the date of the last theoretical dose. For example, if a patient bought a box of 30 pills on January 1^{st}, the event has to occur before March 31 (30+2x30). The date of death is present in the initial data, the switches are identified by the beginning of a new phase of treatment and palliative care as the main diagnosis (which is spotted with a "Z515" tag[4] in the database). Censorship of data caused by the end of the extraction period (end of December 2015) is also considered a legitimate stop. If the data extraction end date is less than two months from the last theoretical consumption, then, in the same way as for legitimate stops, the processing phase is considered censored. If none of the legitimate stops occurs, then the phase is considered to end with an illegitimate stop which means a default of persistence. Of course, this way to look at claims data to observe pharmacy transactions as a way to get a estimation of drug pocession to model non-adherence is perfectible. We plan to improve this with more studies in the future but an analysis of the labeled data did not show obviously wrong labels. The use of explanation methods for "black-box" machine learning algorithms is also a way to detect bias or strange classification results which could underline an error in this process.

3.3 Models

We tested different types of models for our prediction tasks depending on the pre-processing of the data: phases or transaction-centered. In both cases, use a $60-20-20$ train-validation-test splitting at a patient-level (meaning a patient in the train set can't be in the validation or the test ones). For the phase analysis, we compared a logistic regression, a decision tree, a random forest and a gradient boosting to predict, from information available at the beginning of a phase, if it ends with an illegitimate or a legitimate stop. For the transaction-centered study, we fine-tuned the hyper-parameters with optuna (1) to reach the following architecture (visible in Figure 1) :

- a Gate-Recurent Unit (GRU, **(author?)** (Cho *et al.*)) network is applied on the "dynamic" information (pharmacy transactions and hospitalizations)(depth 2, dropout 0.7);

[4]Encounter for palliative care in the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD), a medical classification list by the World Health Organization `https://icd.codes/icd10cm/Z515`

- an MLP network on the "contextual" information (details about the patient (geographical, financial support, ...)) which are not frequently updated in the SNDS database (depth 1, dropout 0.7);
- both outputs are concatenated and then classified through a fully-connected layer (depth 2, dropout 0.7).

"Dynamic" inputs are sequences of 10 observations and the output indicates if the current transaction might not be followed by another one (indicating the risk of an illegitimate drop-out). As 10 observations are not always available for each transaction, we used zeros filling padding after testing different configurations.

3.4 Explanation of the model's decision

Some of the models used are straightforward to explain. For example, the regression estimates a fixed effect of each variable to the patients' average behavior, even if it is based on two strong assumptions (1) the expected effect of each variable is linear and (2) the effect of each variable does not vary over time. Yet, we can use these weights to estimate the influence of a variable on the model's decisions. Other models are more "*black-box*" as it is unclear how they made their decision. The inner working, often based on non-linear decisions, makes it difficult to be humanly understandable and then, they don't provide an estimation of the variable importance on the model decision. To tackle this issue, we surveyed several recent approaches to get insights about the reason behind a given score by a model. After testing different solutions such as Lime (19), we finally use Shap (SHapley Additive exPlanations) (13) as it seems like the advance and efficient solution to explain such model's decisions. In a nutshell, Shap evaluates the contribution of each input feature and uses game theory to get a local explanation of the impact of the features. Each SHAP value expresses the marginal effect that the observed level of a variable for an individual has on the final predicted probability for that individual. Then it can compute an evaluation of the impact of the features on the decision process.

4 Results

First we evaluate the ability of our models to predict a risk of drug drop-out With two measures. First AUC (area under the receiving operating curve) which represents the True-

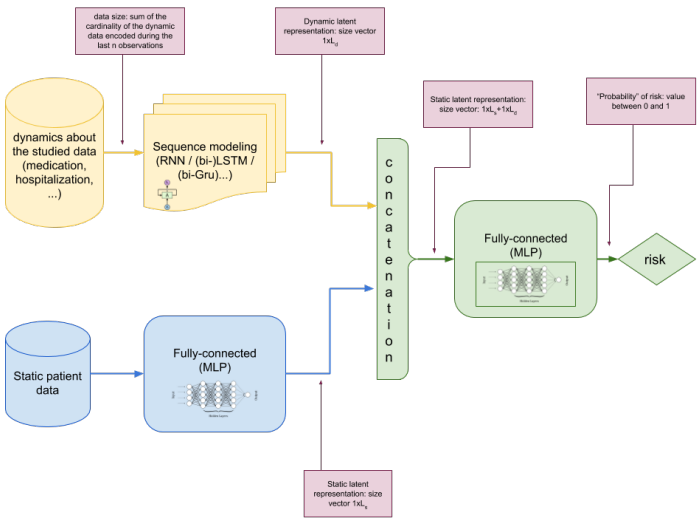
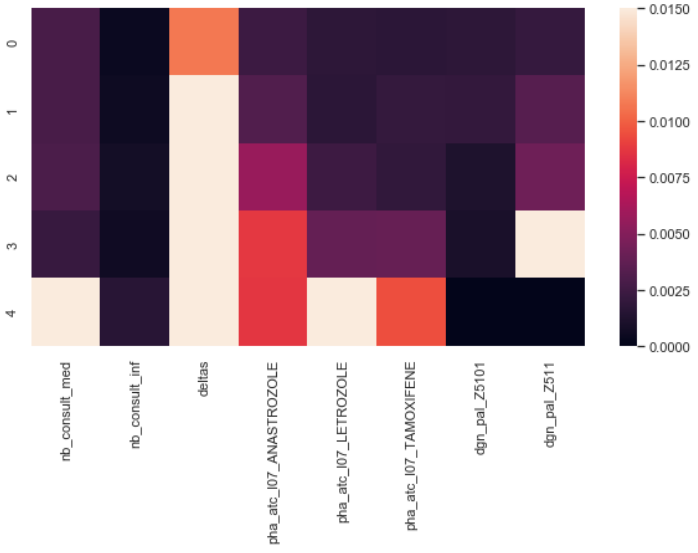


Figure 1: Architecture of our model to analyze patient-centered transactions

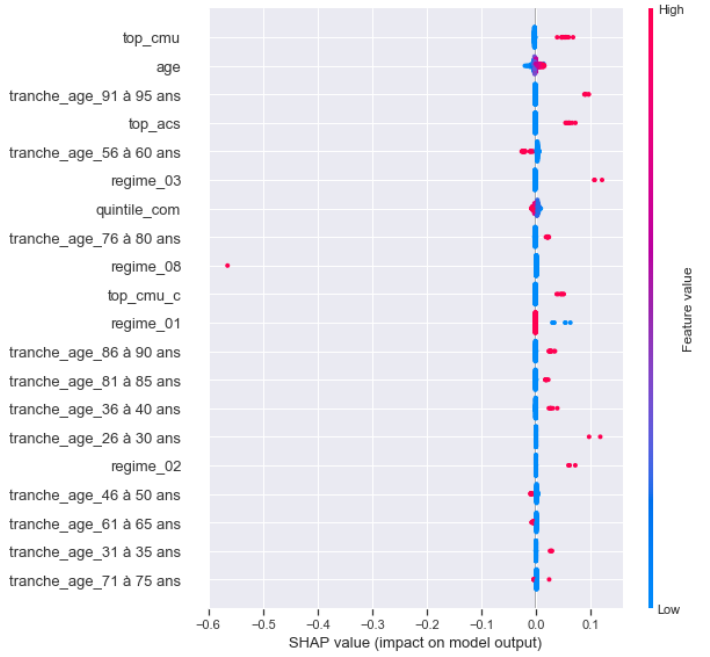
positive rate of the model according to the False-positive rate. It allows to evaluate how well the model is able to discriminate between two classes. Secondly, we use CAP curves (Cumulative accuracy profile) which measure the efficiency of the model to detect true positive. CAP $n\%$ is the rate of true positive classifications looking at the n highest ranked predicted risk.Both phases and transaction-centered models showed their ability to predict an illegitimate drug drop-out with AUCs between 0.68 and 0.74 and the transaction-centered model obtains a 0.82 AUC.

We adapt Shap to to explain the decision of the model both on the dynamic and static features used in our model as shown in Figure 2a and Figure 2b. The figures show how each features is contributing to push the model output from the base value (the average model output over the training dataset we passed). From the dynamic part, we can see that the current drug has a major influence. Then we see the variable *deltas* (linked to the frequency of visit to the pharmacy) and *mpr* (Medication Possession Ratio i.e. how much drug the patient posses compared to the dosage) have a strong influence in the decision process. We also find again that the duration since the declaration of the affection (*log anc ald*) influence also the model which is coherent with the literature. A disappointment is in the exploration of hospitalization, encoded with *dgn pal* labels: results are difficult to interpret and we don't see anything with *dgn pal F 0* line corresponding to psychiatric illness. Yet, our encoding was very naive and an embedding there should improve our results. In the static part, we show the influence of age and also find back the influence of income through *cmu, cmu c and acs* codes that are linked to states supports. A surprising result is an influence of "departement" which are administrative divisions of France. We first try to correlate these weights with the overall population of each department but it doesn't match. We now try to get more detailed information about each district (such as population, wealth, density, . . .).

Finally, we analyze the Shap values computed for each patients as the marginal contribution of each features used by the model to establish its decision. Indeed, each observation gets its own set of SHAP values indicating the weight of each variable into the decision process to move the risk from random. This give insights about the decision process for each case and



(a) Evolution of SHAP value for dynamic features at each steps of our input data (i.e. pharmacy transactions)



(b) Shap values of the static features with their impact on the decision process

Figure 2: Shap evaluation of contribution of features over the decision of the model

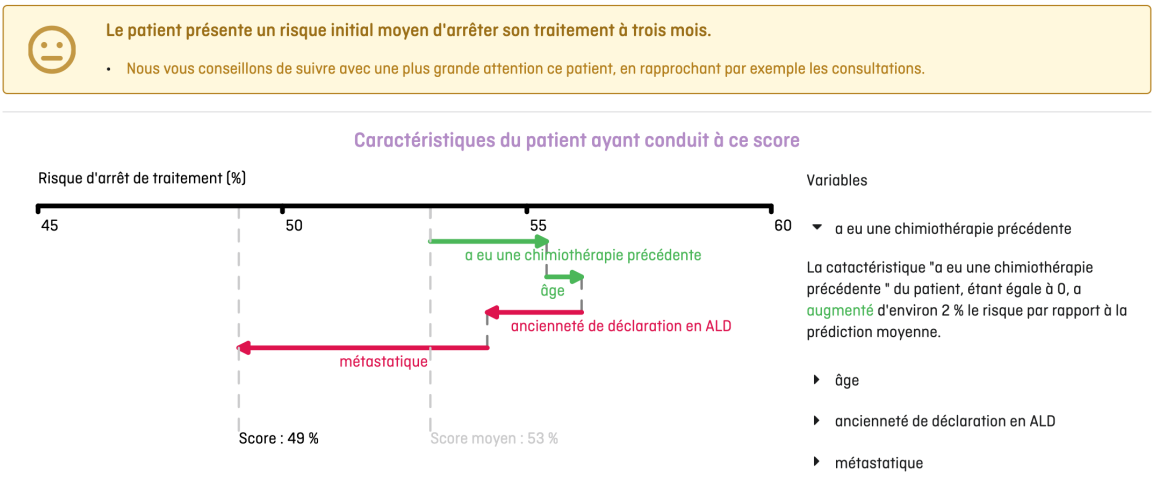


Figure 3: Practical usage of Shap explanation over the decision process for a patient. Green means the variable as a positive effect to lower the risk while red means increase

each feature. A discussion with oncologists validate these results and showed their interest in theses explanations. Yet, the large number of features used in our model and the raw output of Shap algorithms remain difficult to interpret for a novice. We adapt our solution to integrate as simply as possible the notion of the marginal contribution of a feature over the decision process and ended with the visualization in Figure 3.

5 Conclusion and discussion

This paper presents the different steps we followed to build machine-learning methods used to help patients follow their treatment during long-term illness. The focus is set on how we can visualize the data, the models and explain their decision. Working close to the medical staff, we have insight into how important this information is to convince them to trust the decision process.

We validate our first results with feedbacks from oncologists, medical researchers and by comparing them to the literature. We validate the ability of AI to estimate the risk of non-persistence and we explore methods to open these black-boxes to explain the decision taken by the models. The results are coherent with the feedback from oncologists and the literature. Yet some of them are difficult to understand.

These results led us to add more context into the features to improve the model decision while measuring the impact of each kind of new feature. We also plan to extend these approach of other problems such as non-observance (i.e. taking the medication but with a wrong dose) or re-hospitalizations for example.

References

[1] AKIBA T., SANO S., YANASE T., OHTA T. & KOYAMA M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, p. 2623–2631.

[2] CABITZA F., RASOINI R. & GENSINI G. F. (2017). Unintended consequences of machine learning in medicine. volume 318, p. 517–518: American Medical Association.

[Cho *et al.*] CHO K., VAN MERRIENBOER B., GULCEHRE C., BAHDANAU D., BOUGARES F., SCHWENK H. & BENGIO Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[4] DENAXAS S., STENETORP P., RIEDEL S., PIKOULA M., DOBSON R. & HEMINGWAY H. (2018). Application of clinical concept embeddings for heart failure prediction in uk ehr data. *NIPS ML4H: Machine Learning for Health arXiv preprint arXiv:1811.11005*.

[5] DIMATTEO M. R. (2004). Social support and patient adherence to medical treatment: a meta-analysis. volume 23, p. 207: American Psychological Association.

[6] DIMATTEO M. R., LEPPER H. S. & CROGHAN T. W. (2000). Depression is a risk factor for noncompliance with medical treatment: meta-analysis of the effects of anxiety and depression on patient adherence. volume 160, p. 2101–2107: American Medical Association.

[7] FRANKLIN J. M., SHRANK W. H., LII J., KRUMME A. K., MATLIN O. S., BRENNAN T. A. & CHOUDHRY N. K. (2016). Observing versus predicting: Initial patterns of filling predict long-term adherence more accurately than high-dimensional modeling techniques. *Health services research*, **51**(1), 220–239.

[8] HERENT P., SCHMAUCH B., JEHANNO P., DEHAENE O., SAILLARD C., BALLEYGUIER C., ARFI-ROUCHE J. & JÉGOU S. (2019). Detection and characterization of mri breast lesions using deep learning. *Diagnostic and interventional imaging*.

[9] HINTON G. (2018). Deep learning—a technology with the potential to transform health care. volume 320, p. 1101–1102: American Medical Association.

[10] JANSSOONE T., BIC C., KANOUN D., HORNUS P. & RINDER P. (2018). Machine learning on electronic health records: Models and features usages to predict medication non-adherence. *NIPS ML4H: Machine Learning for Health, arXiv preprint arXiv:1811.12234*.

[11] KRUEGER K. P., BERGER B. A. & FELKEY B. (2005). Medication adherence and persistence: a comprehensive review. *Advances in therapy*.

[12] LO-CIGANIC W.-H., DONOHUE J. M., THORPE J. M., PERERA S., THORPE C. T., MARCUM Z. A. & GELLAD W. F. (2015). Using machine learning to examine medication adherence thresholds and risk of hospitalization. volume 53, p. 720: NIH Public Access.

[13] LUNDBERG S. M. & LEE S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, p. 4765–4774.

[14] MANN D. M., WOODWARD M., MUNTNER P., FALZON L. & KRONISH I. (2010). Predictors of nonadherence to statins: a systematic review and meta-analysis. volume 44, p. 1410–1421: SAGE Publications Sage CA: Los Angeles, CA.

[15] MCDONALD H. P., GARG A. X. & HAYNES R. B. (2002). Interventions to enhance patient adherence to medication prescriptions: scientific review. *Jama*, **288**(22), 2868–2879.

[16] MOREL M., BACRY E., GAÏFFAS S., GUILLOUX A. & LEROY F. (2017). Convsccs: convolutional self-controlled case series model for lagged adverse event detection. *Biostatistics (Oxford, England), arXiv preprint arXiv:1712.08243*.

[17] NEUMANN A., WEILL A., RICORDEAU P., FAGOT J., ALLA F. & ALLEMAND H. (2012). Pioglitazone and risk of bladder cancer among diabetic patients in france: a population-based cohort study. *Diabetologia*, **55**(7), 1953–1962.

[18] POVYAKALO A. A., ALBERDI E., STRIGINI L. & AYTON P. (2013). How to discriminate between computer-aided and computer-hindered decisions: a case study in mammography. volume 33, p. 98–107: Sage Publications Sage CA: Los Angeles, CA.

[19] RIBEIRO M. T., SINGH S. & GUESTRIN C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, p. 1135–1144: ACM.

[20] SHICKEL B., LOFTUS T. J., ADHIKARI L., OZRAZGAT-BASLANTI T., BIHORAC A. & RASHIDI P. (2019). Deepsofa: A continuous acuity score for critically ill patients using clinically interpretable deep learning. volume 9, p. 1879: Nature Publishing Group.

[21] TUPPIN P., DE ROQUEFEUIL L., WEILL A., RICORDEAU P. & MERLIÈRE Y. (2010). French national health insurance information system and the permanent beneficiaries sample. *Revue d'epidemiologie et de sante publique*, **58**(4), 286–290.

Exploitation de documents médicaux par les techniques d'embedding : application au typage automatique de documents

Mikaël Dusenne[1,2], Julien Grosjean[1,2], Lina Soualmia[2,3], Clément Massonnaud[1], Stéphane Canu[3], Stefan J. Darmoni[1,2]

[1] DÉPARTEMENT D'INFORMATION ET D'INFORMATIQUE BIOMÉDICALE, Centre Hospitalo-Universitaire de Rouen, Rouen, France
mikael.dusenne@chu-rouen.fr

[2] UMR_S 1142, F-75006 Sorbonne Univ., Inserm LIMICS, Paris, France

[3] LITIS EA 4108, F-76000, Normandie Univ., UNIROUEN, UNIHAVRE, INSA Rouen, Rouen, France

Résumé : **Introduction :** Les documents non structurés contiennent la majeure partie de l'information utile d'un dossier patient informatisé. Les techniques de traitement automatique de la langue permettant d'exploiter ces données sont en constante évolution. Les techniques d'embedding transforment des concepts non structurés en un espace vectoriel multidimensionnel. Il est ensuite possible d'exploiter les données sous forme numérique afin d'accomplir différentes tâches, supervisées ou non (classification, création de clusters sémantiques, quantification de similarité sémantique).
Méthodes : Nous étudions dans cet article une des possibles applications des techniques d'embeddings aux documents médicaux non structurés. En utilisant plus de 15 millions de documents médicaux disponibles dans l'entrepôt de données cliniques du Centre Hospitalo-Universitaire de Rouen, nous créons des *document embeddings* avec doc2vec après avoir identifié les hyper-paramètres optimaux sur un sous-ensemble de documents. Un réseau de neurones est ensuite entraîné pour prédire le type de document en fonction de l'embedding qui le représente. Nous analysons les performances de classification en utilisant le pourcentage de bonne classification.
Résultats : Le choix des hyper-paramètres fait grandement varier la qualité des embeddings générés, et ces paramètres semblent être très dépendants des données utilisées. La classification des types de documents présentait un pourcentage de classification correcte de 99,07% sur 110 481 documents de l'ensemble de test.
Conclusion : Les premiers résultats obtenus montrent que les techniques d'*embedding* semblent offrir des avantages supérieurs aux autres méthodes de traitement automatiques de la langue, et permettent de répondre à des problématiques nouvelles.
Mots-clés : Apprentissage automatique, Apprentissage profond, Réseaux de neurones, Traitement automatique de la langue

1 Introduction

L'exploitation de données produites par les hôpitaux peut être très utile dans un contexte de recherche de médicale et de construction d'outils d'aide à la prise en charge des patients. Les données pertinentes d'un entrepôt de données de santé (EDS) sont représentées à 80% sous forme non structurée. (Raghavan *et al.*, 2014) ont par ailleurs montré que les données non structurées étaient essentielles pour répondre aux critères d'inclusion des études cliniques dans 59% à 77% des cas. Afin de fournir des résultats intéressants face à diverses problématiques, il est nécessaire d'exploiter ces données de la façon la plus efficace possible.

Les techniques de traitement automatique de la langue (TAL) existent depuis de nombreuses années et abordent la tâche de différentes manières. Le principal inconvénient des techniques classiques est leur représentation peu efficace du texte, dont la très haute dimensionnalité résulte en une sparcité importante, et dont la nature ne permet pas de convoyer les liens sémantiques qui existent entre les mots d'un texte.

L'absence de technique permettant de représenter des mots sous forme numérique de façon efficace a été un frein majeur au développement du TAL.

Les techniques de *word embedding* (Mikolov *et al.*, 2013) consistent en un apprentissage semi supervisé de données non structurées, et transformant chaque mot en un vecteur de

nombres réels. Ces vecteurs ont une dimension ne dépendant pas de la taille du vocabulaire, et seront généralement bien plus denses que les données utilisées dans les algorithmes de TAL reposant sur le concept de sac de mots. De plus, les vecteurs générés conservent une valeur sémantique, et les mots sémantiquement semblables seront proches les uns des autres dans l'espace vectoriel. Cette dernière propriété est absente des algorithmes classiques, et offre la possibilité d'effectuer, par le biais d'opérations vectorielles, des transformations et rapprochements sémantiques, permettant de réaliser à la fois des tâches de classification, supervisées, et des tâches non supervisées telles que la création de clusters sémantiques.

Les implémentations des *word embeddings* sont nombreuses. Word2vec (Mikolov *et al.*, 2013), développée en 2013, est la plus connue. Elle était la première implémentation facilement utilisable et fournissant des résultats surpassant l'état de l'art sur les tâches évaluées.

Depuis, de nombreuses autres implémentations ont été créées, abordant la façon de générer les vecteurs de façons variées. De plus, la notion d'embedding peut être considérée de façon plus abstraite et par exemple doc2vec (Le & Mikolov, 2014), développé en 2014, permet de générer un vecteur par document (et non pas par mot). Cela permet donc d'appliquer les techniques d'embeddings et de regrouper des documents en fonction de leurs contenus sémantiques, afin de les catégoriser automatiquement, ou de retrouver les documents les plus proches d'un document donné.

(Dynomant *et al.*, 2019a) ont exploré doc2vec dans un contexte de littérature médicale, utilisant les résumés des articles issus de PubMed afin d'implémenter une fonctionnalité de recommandation d'articles similaires à un article donné. Bien qu'utilisant une ressource en langue anglaise et n'étant pas composée de documents médicaux, les résultats montrent que Doc2vec est un outil intéressant et mérite d'être exploité dans le contexte d'un EDS. Ces outils offrent des possiblités d'exploitation des données non structurées jusqu'ici impossibles à mettre en place, comme la catégorisation automatique de documents.

Les documents de santé peuvent être de différents types, parmi lesquels figurent : compte-rendu d'hospitalisation, compte-rendu d'acte, compte-rendu d'accouchement, ordonnance, compte-rendu de biologie. Ces types sont renseignés dans le système d'information manuellement lors de la saisie du document. Cependant, de nombreux documents du système d'information du CHU de Rouen ne sont pas typés. Le développement d'un outil capable de déterminer automatiquement le type d'un document permettrait d'améliorer la qualité des données du système d'information et de l'EDS, et d'éviter la saisie manuelle du type de futurs documents. Cette tâche supervisée est aussi un moyen d'évaluer objectivement la capacité des *embeddings* à exploiter des données non structurées pour accomplir une tâche concrète.

L'objectif de ce travail est d'explorer l'utilisation des *document embeddings* pour catégoriser automatiquement les différents types de documents médicaux non structurés issus d'un entrepôt de données de santé.

2 Méthodes

Le Centre Hospitalo-Universitaire (CHU) de Rouen dispose d'un entrepôt de données médicales contenant 15,7 millions de documents médicaux concernant deux millions de patients, issus de l'activité de l'hôpital entre 1992 et 2019.

Ces documents sont constitués entre autres de comptes-rendus hospitaliers, comptes-rendus de consultation, ordonnances, comptes-rendus d'actes, de chimiothérapie, de résultats de laboratoire, rédigés en langue française.

Notre travail, s'inscrivant dans le cadre d'une thèse de science débutée en Octobre 2019 et faisant suite à celle réalisée par Émeric Dynomant, consiste en l'étude de l'application des *embeddings* à plusieurs niveaux d'agrégation :

— *Word Embeddings* : création d'un annotateur sémantique hybride, utilisant l'annotateur sémantique déjà existant dans l'entrepôt (Sakji *et al.*, 2010), basé sur les techniques de sacs de mots. L'objectif est d'améliorer les performances de l'annotateur notamment au niveau des erreurs d'orthographes, nombreuses dans les documents, et des abréviations, nombreuses elles aussi et dont le sens dépend fortement du contexte (une même abréviation peut avoir un sens différent selon la spécialité).

— *Document Embeddings* : création de embeddings centrés sur les documents, permettant de classer automatiquement les différents types de documents (par exemple compte-rendu d'hospitalisation, compte-rendu d'acte, ordonnance, compte-rendu de biologie). La détermination du type de document permettra de compléter cette donnée manquante pour 4,6 millions de documents (16%) de l'EDS et donc d'améliorer la qualité des données.
— *Séjour embeddings* : l'agrégation des documents par séjour permettrait d'aider à la codification de la tarification à l'acte de l'hôpital grâce à l'implémentation d'un classifieur basé sur le diagnostic principal de la terminologie CIM-10, et des diagnostics secondaires.
— *Patient Embeddings* : la création d'un vecteur pour chaque patient permettrait de rapprocher automatiquement les patients ayant des antécédents médicaux similaires. Cela pourrait permettre d'implémenter des outils d'aide à l'inclusion dans des cohortes pour la recherche médicale, mais aussi des outils d'aide à la prise en charge en retrouvant les patients similaires à un patient donné.

La première tâche que nous avons réalisée et dont nous présentons les résultats dans cet article est la classification du type de document grâce aux *document embeddings*.

Les 15,6 millions documents de L'EDS du CHU de Rouen sont de dix types distinct. En dehors des 4,62 millions de documents non typés, on compte 5,61 millions de compte-rendus d'acte, 3,06 millions ordonnances, 2,02 millions compte-rendus de séjour, 1,97 millions compte-rendus opératoires, 72 209 compte-rendus de chimiothérapie, 41 879 compte-rendus de consultation, 25 917 compte-rendus de biologie, 19 525 documents à visée légale, 8 652 compte-rendus de soins intensifs post-opératoires, 2 783 compte-rendus d'accouchement.

L'utilisation de doc2vec nous permet, après entraînement sur le corpus, d'obtenir un vecteur pour chaque document. Il existe de nombreux paramètres d'entraînements, et leurs valeurs optimales dépendent largement du corpus utilisé et ne peuvent pas être déterminées à l'avance. Afin d'obtenir un espace vectoriel de qualité optimale, nous avons utilisé un premier ensemble de 100 000 documents pour effectuer une optimisation de dix hyper-paramètres principaux. Pour chaque combinaison de paramètres, nous avons effectué une tâche de classification basée sur la méthode des "K plus proches voisins" (classification Kppv). Ce type de classification présente l'avantage de ne reposer que sur la distance entre les points de l'espace vectoriel et exploite donc les vecteurs produits sans ajouter un classifieur qui nécessiterait un entraînement et ajouterait une complexité non désirée à cette étape de l'analyse. La qualité des vecteurs était donc évaluée par le taux de bonne classification du type de document.

Après avoir obtenu les meilleurs paramètres (la combinaison offrant les meilleures performances de classification par Kppv), nous avons séparé les documents restants en un ensemble d'entraînement des *embeddings* (90%), un ensemble d'entraînement du classifieur (9%) et un ensemble de validation (1%). Les *embeddings* nécessitent un large corpus pour être capable d'apprendre la meilleure représentation des documents, simplifiant au maximum la classification, ce qui explique l'allocation de 90% du corpus dans cette étape.

Le réseau de neurones du classifieur est implémenté avec la bibliothèque python keras [1]. Sa structure est gardée volontairement simple, car les vecteurs obtenus après entraînement devraient permettre une classification facile ne nécessitant pas un classifieur complexe. Pour la validation de l'entraînement, 10% des documents dédiés à l'entraînement de ce réseau de neurones seront utilisés.

Les calculs ont été réalisés sur un serveur hébergé au CHU de Rouen, disposant de 194 cœurs et 1To de RAM.

1. `https://keras.io/`

3 Résultats

3.1 Optimisation des hyper-paramètres

L'optimisation des hyper-paramètres a révélé une grande sensibilité des performances à certains paramètres d'entraînement. Notamment, l'algorithme PV-DBOW (Distributed Bag Of Words version of Paragraphs Vectors) donnait de meilleurs résultats que PV-DM (Distributed Memory version of Paragraphs Vectors), avec une exactitude de 0,85 (SD=0,11) et 0,68 (SD=0.07) respectivement (test de student apparié : p-value $<$ 0,0001). Ce résultat, allant à l'encontre des résultats retrouvés initialement par (Le & Mikolov, 2014), montre l'importance de ce travail d'optimisation pour chaque jeu de données. Le travail de (Dynomant *et al.*, 2019a) retrouvait les mêmes conclusions sur les meilleures performances de PV-DBOW.

L'augmentation du nombre d'époques permettait d'améliorer les performances de manière fiable, et nous n'avons pas retrouvé de surapprentissage.

Le meilleur modèle utilisait PV-DBOW et des vecteurs de 200 dimensions. Les autres paramètres étaient : window=8, negative=20, hs=0, min_alpha=0,0025, min_count=20, sample=0.

Étant donné que le nombre d'époques montrait une hausse des performances de façon fiable, et étant donné que le nombre de documents du jeu d'entraînement est beaucoup plus important, nous avons décidé d'entraîner le modèle final avec 200 époques. Les autres paramètres sont disponibles en annexe.

3.2 Entraînement des *embeddings*

L'entraînement a duré 6 jours et 7 heures.

Afin de pouvoir visualiser les *embeddings* en 3 dimensions, nous avons réalisé une réduction de dimensionnalité par T-SNE, une méthode de projection permettant de réduire le nombre de dimensions tout en conservant au mieux les distances entre les différents points d'un espace vectoriel.

Cette étape ne fait pas partie du processus de classification des documents, mais la visualisation des *embeddings* peut nous donner des indices sur leur structure. Afin de limiter le temps de calcul, nous avons sélectionné un sous-ensemble de 300 000 documents de façon aléatoire, et calculé la projection en trois dimensions de nos vecteurs de 200 dimensions, puis coloré chaque document selon son type (**Figure 1**).

On constate que les types de documents sont bien séparés en différents clusters, et donc que doc2vec semble avoir été capable d'identifier de façon automatique ces différents types de documents.

3.3 Entraînement du classifieur

Le réseau de neurones utilisé est constitué de deux couches denses de 32 unités avec une fonction d'activation ReLu (Rectified Linear unit), chacune suivie d'une couche de dropout à 29%. La dernière couche est une couche dense de 10 unités avec activation softmax permettant d'obtenir le type de document prédit, et a été entraîné pendant 200 époques. Il prend en entrée les différents vecteurs générés par doc2vec, puis retourne les types de documents prédits.

Il n'y avait pas de signe de surentraînement (pas de diminution du taux de bonne réponse (exactitude) de l'ensemble de validation), et l'évolution de l'exactitude était asymptotique et stabilisée à la fin de l'entraînement du réseau de neurones (**Figure 2**).

La visualisation des *embeddings* par T-SNE nous permettait de supposer qu'un classifieur simple devrait parvenir à exploiter les vecteurs facilement. En effet, nous n'avons pas obtenu de meilleurs résultats en augmentant le nombre de couches ou d'unités, et avons conservé le modèle le plus simple qui exploitait au mieux l'information des *document embeddings*.

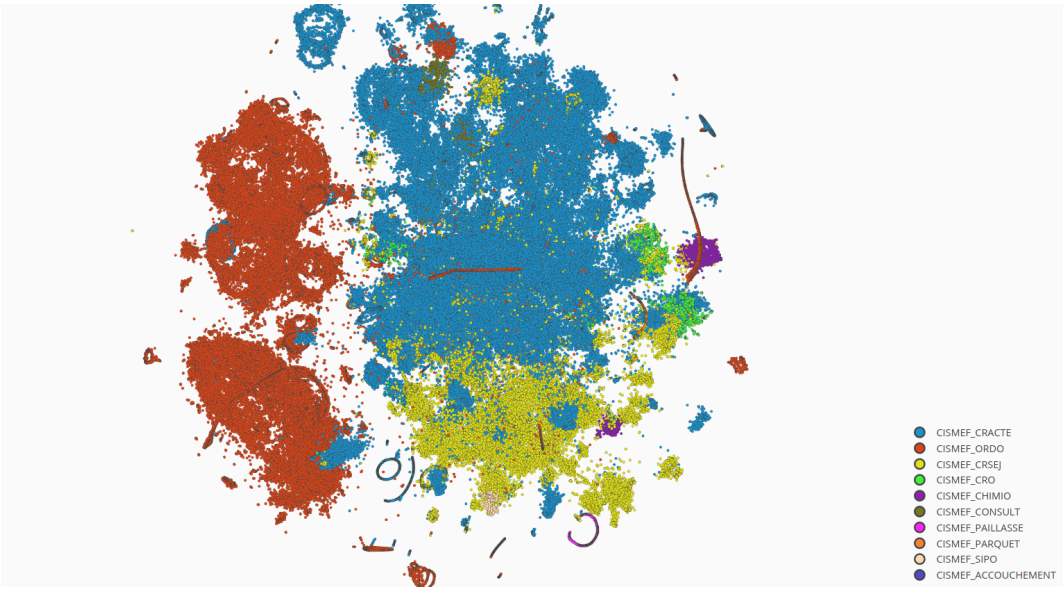


FIGURE 1 – Représentation 3D T-SNE des vecteurs de documents obtenus. Les couleurs superposées représentent les types de documents. On peut observer que les documents de même type se trouvent proches dans l'espace vectoriel, ce qui nous indique que doc2vec semble avoir été capable de séparer les documents de façon autonome.

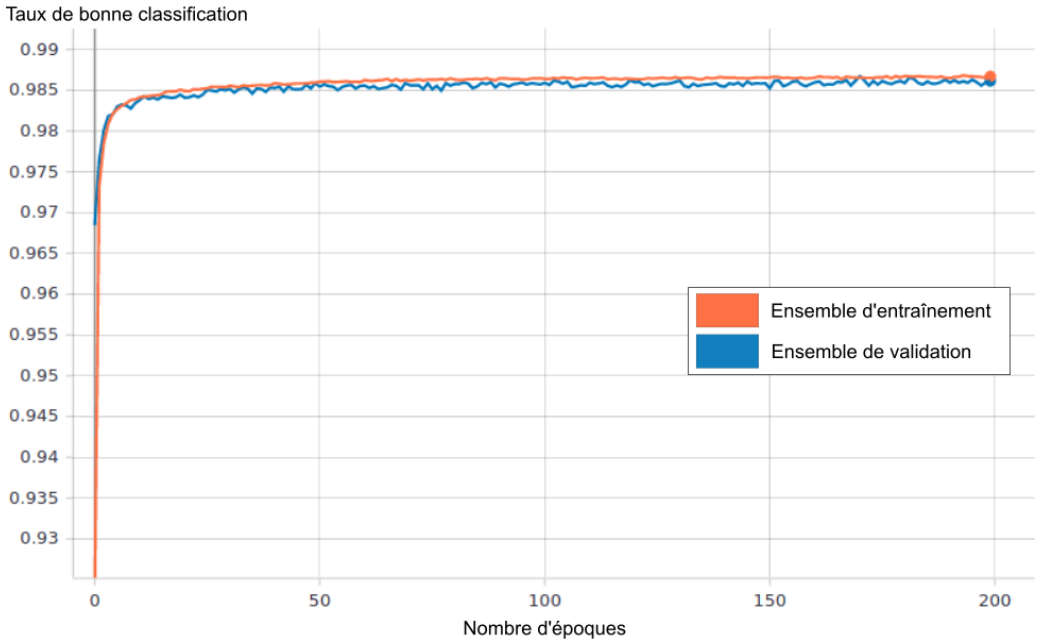


FIGURE 2 – Évolution de l'exactitude de l'ensemble d'entraînement et de validation, en orange et en bleu respectivement. En abscisse figure le nombre d'époques écoulées, en ordonnée l'exactitude.

3.4 Performances de classification

Le taux de bonne classification était de 99,07% sur les 110 481 documents évalués.
Le pourcentage de bonne classification était variable selon le type réel du document.

L'exactitude moyenne était de 97,54% (SD=2,35%) avec un minimum de 95,31% (61/64) pour les documents de soins intensifs post-opératoires, et un maximum de 100% (241/241) pour les documents issus des laboratoires de biologie.

Les principales erreurs étaient la prédiction "compte-rendu d'acte" alors qu'il s'agissait d'un compte-rendu de séjour, et inversement. La matrice de confusion des classifications est représentée sur la **Figure 3**.

Exactitude: 99.07 % (110481 Documents)

pred \ actual	ACCOUCHEMENT	CHIMIO	CONSULT	CRACTE	CRO	CRSEJ	ORDO	PAILLASSE	PARQUET	SIPO
ACCOUCHEMENT	25	0	0	0	0	0	0	0	0	0
CHIMIO	0	705	0	0	0	14	0	0	0	0
CONSULT	0	0	353	59	0	0	0	0	0	0
CRACTE	0	0	30	56001	36	162	31	0	1	0
CRO	0	0	0	10	1967	2	0	0	0	0
CRSEJ	1	24	0	601	4	19643	2	0	0	3
ORDO	0	0	0	36	0	2	30277	0	0	0
PAILLASSE	0	0	0	0	0	0	0	241	0	0
PARQUET	0	0	0	4	0	0	0	0	178	0
SIPO	0	0	0	0	0	8	0	0	0	61
Exactitude par type	96.15	96.71	92.17	98.75	98.01	99.05	99.89	100	99.44	95.31

FIGURE 3 – Table de confusion de la classification des documents de l'ensemble de validation. En colonne figurent les types de références dans l'entrepôt, en ligne les types prédits par le réseau de neurones.

Afin d'explorer les erreurs de classification, il est possible d'utiliser le niveau de confiance du réseau de neurones. Si le réseau de neurones fait plus d'erreurs de classification lorsqu'il présente un faible niveau de confiance dans le type prédit, l'exactitude sera plus élevée si on ne garde que les documents pour lesquels la confiance de prédiction était élevée.

Le niveau de confiance peut être approximé en utilisant les composantes de *dropout* du réseau de neurones (Gal *et al.*, 2016) : en effet, ces couches mettent à zéro une fraction de leurs neurones à chaque document fourni en entrée, de façon aléatoire. Elles sont habituellement utilisées afin de réduire le surapprentissage en servant de couche de régularisation, et dans ce cadre elles sont inactivées lorsque que l'entraînement est terminé.

Mais il est possible de garder ces couches actives après l'entraînement et d'introduire de ce fait une composante non déterministe dans les prédictions réalisées. Si l'on répète un grand nombre de fois une même prédiction, on peut agréger les résultats obtenus et calculer un indice de confiance (proportion de la classe prédite majoritaire / nombre de tentatives).

La **Figure 4** représente l'évolution de l'exactitude et du nombre de documents restants en fonction du seuil minimum de confiance jugé acceptable. Le niveau de confiance moyen était de 0,993 (DS = 0,039).

La confiance était en moyenne très élevée, et il y a très peu de documents ayant une confiance de prédiction inférieure à 98%. On constate cependant que l'exactitude augmente de 99,06% à 99,60% avec un seuil de confiance minimal fixé à 98%, et en ayant éliminé 7 326 (6,6%) documents ayant une confiance inférieure.

On peut cependant en conclure que pour de nombreux documents mal catégorisés, l'indice de confiance était élevé.

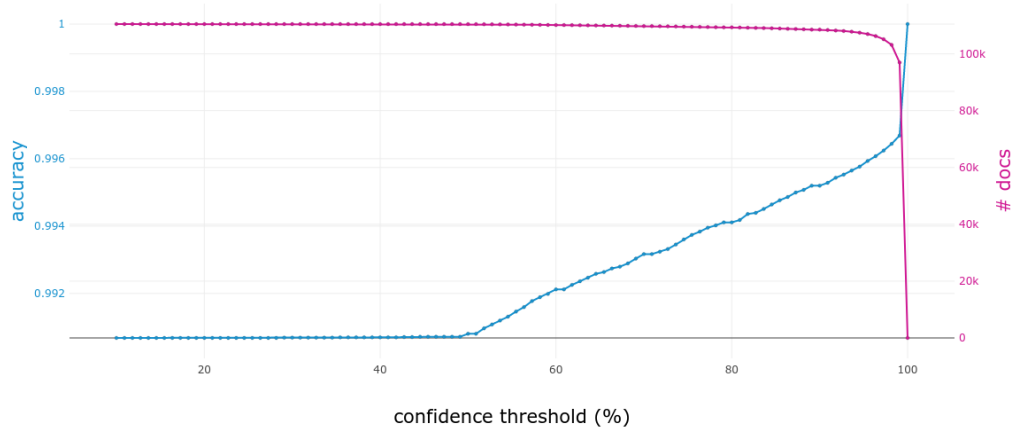


FIGURE 4 – Graphique de rejet du classifieur. En abscisse figure la certitude du réseau de neurones (de 0% à 100%) minimale pour calculer l'exactitude. En ordonnée en bleu (axe de gauche, représenté entre 99% et 100% d'exactitude pour la lisibilité de la figure) figure l'exactitude correspondante, et en rose (axe de droite) figure le nombre de documents considérés pour le calcul de l'exactitude. On constate que la confiance du modèle dans ses prédictions était forte, et il y a très peu de documents exclus avant le seuil de 98%. Au seuil de 99% de confiance, l'exactitude n'a que légèrement augmenté, indiquant que, lorsque le réseau était peu confiant, il se trompait plus fréquemment, mais que la plupart du temps le classifieur était certain de ses résultats.

4 Discussion

Les résultats de l'optimisation des hyper-paramètres pour l'entraînement de doc2vec nous permettent de constater l'importance de cette étape lors de l'utilisation des *embeddings* pour effectuer du traitement automatique de la langue.

Après avoir identifié les meilleurs paramètres pour notre jeu de données, nous avons pu obtenir de bons résultats de classification, avec un taux d'erreur inférieur à 1%, avec un classifieur relativement peu complexe.

Ceci confirme le fait que doc2vec est capable de créer des *embeddings* ayant des propriétés intéressantes et exploitables pour répondre à des problématiques concrètes.

Les erreurs de classifications existaient principalement entre compte-rendu d'acte et compte-rendu d'hospitalisation. Cela pourrait être causé par le lien fréquent qui existe entre ces documents, en effet, de nombreux patients ayant été hospitalisés ont aussi bénéficié d'un acte médical, et le déroulement de cet acte est repris dans le compte-rendu d'hospitalisation.

Les erreurs de typage peuvent être partiellement corrigées en excluant les documents pour lesquels la certitude du classifieur est plus faible. Cela n'empêche pas une mise en application pratique dans laquelle les documents non catégorisés seraient soumis à un évaluateur humain pour une classification manuelle, la majorité des documents étant traités automatiquement.

Pour les documents mal catégorisés malgré une grande certitude, on ne peut exclure la possibilité d'un mauvais typage manuel du document lors de sa création, et la suite de ce travail inclut une évaluation humaine des documents mal classifiés afin de déterminer précisément la cause de ces erreurs. Si ces erreurs sont avérées, notre outil pourrait ainsi être utilisé pour corriger les erreurs existantes dans l'EDS, et donc améliorer encore la qualité des données.

Ce premier travail nous conforte dans l'idée que cette technologie est capable d'apprendre sur des documents médicaux rédigés en français, et nous permet de continuer à l'explorer pour des tâches a priori plus complexes.

5 Conclusion et perspectives

Dans cet article nous avons étudié de façon approfondie une application des *embeddings* aux documents textuels médicaux, la classification des types de documents.

Nous avons démontré que les *embeddings* peuvent être utilisés sur des documents médicaux rédigés en français, et être utilisés avec une bonne fiabilité, qui sera confirmée par une évaluation manuelle d'un échantillon de documents.

Les autres applications des *embeddings* qui pourraient répondre à des problématiques de la recherche en santé, citées au début de cet article, seront implémentées lors de travaux futurs.

L'implémentation de l'annotateur hybride débutera par la création de *word embeddings*.

Pour la réalisation de ce travail nous nous baserons sur les résultats d'une étude précedemment réalisée au CHU de Rouen (Dynomant *et al.*, 2019b) , qui comparait word2vec, GloVe et Fasttext, des implémentation de *word embeddings* utilisant des algorithmes différents, sur un sous-ensemble des documents médicaux de l'EDS. Les résultats de cette étude montraient que word2vec avec l'architecture skip-gram présentait les meilleures performances, et nous débuterons l'implémentation en utilisant cette solution.

Cette étape nous permettra de plus d'explorer des implémentations récentes de *word embeddings*, notamment ALBERT (Lan *et al.*, 2019), une amélioration de BERT (Devlin *et al.*, 2018) développée en décembre 2019 par Google. Biobert (Lee *et al.*, 2019) est un modèle pré-entraîné à la fois sur un corpus généraliste (Wikipedia) et sur un corpus biomédical (abstracts de pubmed et texte des articles de PMC). Les performances de ce modèle sont intéressantes, mais le corpus d'entraînement n'étant pas en langue française, il ne peut pas être utilisé dans notre cas. Il sera intéressant d'utiliser des modèles pré-entraînés sur des corpus en langue française tels que camemBERT (Martin *et al.*, 2019). Ce modèle a été entraîné sur un très large corpus en français, sans orientation médicale. Il sera intéressant de comparer les performances à celles d'un modèle entraîné exclusivement sur des documents issus de l'EDS de Rouen.

La création de *Séjour embedding* sera différente du cas décrit dans cet article selon deux modalités : premièrement, un séjour est constitué d'un ensemble de documents médicaux produits au cours d'une hospitalisation. Il sera donc nécessaire de procéder à une agrégation afin de produire un vecteur par séjour. Une moyenne des différents vecteurs de documents est envisageable, mais il est aussi possible de concaténer les documents en amont afin d'obtenir un vecteur unique. Enfin, la variable à prédire est représentée par les codes de diagnostique de la CIM-10 : il existe donc un très grand nombre de modalités (environ 16 000). Ce facteur nécessitera de repenser la façon dont le classifieur est implémenté. La structure en arbre de la terminologie est un des éléments qui sera mis à contribution afin de rendre cette classification plus efficace.

Concernant la création de *patient embeddings*, il sera nécessaire de prendre en compte la dimension temporelle des différents séjours réalisés à l'hôpital pour ces patients. Une des pistes est l'utilisation de réseaux de neurones récurrents du type LSTM (Long Term - Short Term Memory).

Références

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding.

DYNOMANT E., DARMONI S. J., ÉMELINE LEJEUNE, KERDELHUÉ G., LEROY J.-P., LEQUERTIER V., CANU S. & GROSJEAN J. (2019a). Doc2vec on the pubmed corpus : study of a new approach to generate related articles.

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

GAL Y., YG279@CAM.AC.UK, GHAHRAMANI Z., ZG201@CAM.AC.UK & OF CAMBRIDGE U. (2016). Dropout as a bayesian approximation : Representing model uncertainty in deep learning.

LAN Z., CHEN M., GOODMAN S., GIMPEL K., SHARMA P. & SORICUT R. (2019). Albert : A lite bert for self-supervised learning of language representations.

LE Q. V. & MIKOLOV T. (2014). Distributed representations of sentences and documents. doc2vec.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2019). Biobert : pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv :1901.08746*.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., ÉRIC VILLEMONTE DE LA CLERGERIE, SEDDAH D. & SAGOT B. (2019). Camembert : a tasty french language model.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space.

RAGHAVAN P., CHEN J. L., FOSLER-LUSSIER E. & LAI A. M. (2014). How essential are unstructured clinical narratives and information fusion to clinical trial recruitment ? *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, **2014**, 218–223.

SAKJI S., GICQUEL Q., PEREIRA S., KERGOURLAY I., PROUX D., DARMONI S. & METZGER M.-H. (2010). Evaluation of a french medical multi-terminology indexer for the manual annotation of natural language medical reports of healthcare-associated infections. *Studies in health technology and informatics*, **160**, 252–256.

Graph clustering for hospital communities

T Ngo[1,2], V Georgescu[1], C Gervet[3], A Laurent[2], T Libourel[3], G Mercier[1]

[1] Economic Evaluation Unit, CHU Montpellier, Montpellier, France
[2] LIRMM, Univ Montpellier, CNRS, Montpellier, France
huu-tu.ngo@lirmm.fr

[3] Espace-Dev, Univ Montpellier, IRD, Univ Réunion, Univ Guyane, Montpellier, France

Résumé :
Patients frequently change hospitals, especially for the management of chronic diseases. To ensure efficient and high-quality treatments, doctors would need access to the patients' medical records at the previous hospitals. Therefore, health authorities are interested in building hospital communities whereby medical records can be shared among the hospitals. In this paper, we propose a graph-based approach to address this problem, making use of data from the French national hospital discharge database (PMSI). Particularly, we first model data flows of patients between hospitals as an undirected weighted graph in which nodes and edges present the hospitals and the amount of patient flows respectively. Then, after evaluating two common graph clustering methods, we customize the more suitable one for our needs. Our result is a partition of French hospitals into 19 communities, among which 17 communities are in metropolitan France, and provides interesting insights compared with approaches based on administrative boundaries.

Mots-clés : Graph clustering, Spectral clustering, Louvain, hospital communities, PMSI

1 Introduction

It is a noticeable fact that patients do not visit the same hospitals every time. There are many reasons for that. For example, patients have changed addresses, they are not happy with the service of the previous hospital, or they need to seek specialized care in a tertiary hospital. In such cases, it is clear that the treatment would be more efficient and the risk to patients' health could be eliminated or reduced if the later hospitals were able to access the medical records of the patients at the previous hospitals. In other words, there is a need to allow information technology systems to share medical records among hospitals. However, it is neither necessary nor practical for all hospitals in France to be grouped as one because it would be costly while some hospitals will never share any patient. Therefore, health authorities are interested in building hospital communities so that medical records can be shared among the hospitals in those communities.

In the meantime, in the French context, public hospitals are already grouped into regional hospital groups (fr. Groupements hospitaliers de territoire - GHT). As these GHTs are proposed by the regional health agencies (Agences régionales de santé - ARS), these GHTs have limitations due to the administrative boundaries. In addition, in these GHT, private hospitals are not included. Therefore, a scientific approach at the national level for all hospitals types is of high interest to hospitals, health authorities as well as health scientific communities.

On the other side, in France, a national hospital discharge database (fr. Programme de Médicalisation des Systèmes d'Information - PMSI) is available [1]. This PMSI database stores discharge data from all French public and private hospitals. In particular, this database contains a record for each acute inpatient stay [2] (Boudemaghe & Belhadj, 2017). In other words, the patients' pathway can be described. For example, a patient P has the pathway such as $h1 \rightarrow h2 \rightarrow h2 \rightarrow h1 \rightarrow h2$ in which h1 and h2 are the hospitals the patient has gone to.

To group hospitals into communities we could use graph clustering methods. Particularly, in our approach, patients' flows between hospitals are represented by an undirected graph

1. Upon registration with and payment to a habilitated provider, or through collaboration with a French university hospital health information management department

2. There are about 25 million records per year

in which the nodes represent hospitals and the edges represent the size of patient flows. For example, the pathway of patient *P* above would be plus 3 (2 for h1 - h2 and 1 for h2 - h1) for the edge between h1 and h2 on the undirected graph.

Based on the undirected graph, the goal of our work is to group hospitals into communities. To achieve this goal, two different graph clustering methods, spectral clustering and Louvain in particular, are implemented and their performances in terms of quality on our dataset are compared. Particularly, multi criteria are used to evaluate the performances. The paper is organized as follows. Section 2 briefly introduces the spectral clustering method and Louvain method as well as the dataset and the evaluation method we use for the experiments. Section 3 presents the experimental results. Section 4 concludes the paper.

2 Dataset, Graph Clustering Methods and Evaluation

2.1 Dataset

As mentioned in the introduction section, our work is based on the PMSI database system which keeps record of every hospitalization of any patient at both public and private hospitals. This database system allows us to extract the flows of patients between hospitals. In particular, the patient flows of hospitalizations in three continuous years, 2016 to 2018, are extracted. This dataset contains a total of 1,777 hospitals, either public or private, in France. Among these hospitals, the total number of times patients changed hospitals is 13,094,068. Other descriptive information of the dataset is provided below (Table 1).

TABLE *1 – Descriptive information of the graph presenting patient flow*

Number of nodes	1,777
Number of edges	290,707
Max value of weights	34,248
Min value of weights	1
Median value of weights	2
Total weight	13,094,068

2.2 Graph clustering methods

Graph clustering is also known as graph partitioning or community detection and has been studied and applied in many domains including social network (Souza *et al.*, 2013), chemical informatics (Lam & Chan, 2008), computer vision (Shi & Malik, 2000). Many graph clustering methods have been proposed, including random walk based methods (Harel & Koren, 2001), spectral clustering (Von Luxburg, 2007; Hamad & Biel, 2008), modularity based methods (Vincent *et al.*, 2008). In our work, we consider two approaches which are the spectral clustering method and a modularity-based or Louvain method in paricular. In this section, we briefly present the details of these two methods.

2.2.1 Graph notations

A graph can be presented as G = (V, E) where V = $\{v_1, v_2, ..., v_n\}$, is a set of nodes or vertices and E is a set of edges which are two-element subsets of V like $\{v_i, v_j\}$, with v_i, v_j $\in$ V. In the case of weighted graph, each edge carries a non-negative weight $w_{ij} > 0$.

A matrix W = (w_{ij}) where i, j = 1, ...,n is called *weight matrix*. If the graph is an undirect graph then $w_{ij} = w_{ji}$. Moreover, a *degree matrix D* is a diagonal matrix of $d_1, ..., d_n$ in which d_i is the degree of node $v_i \in$ V computed by :

$$d_i = \sum_{j=1}^{n} w_{ij}$$

Moreover, graph clustering is a process of partitioning a graph into sub graphs. Mathematically, if we split the graph G above into K sub graphs whose sets of the nodes are $A_1, ..., A_K$, then we have $A_1 \cup A_2 \cup ...A_K$ = V and $A_i \cap A_j = \emptyset$ with any $i \neq j$ and i, j = 1, ..., K. To measure the qualities of the graph clustering, we define :

— The total weights to be lost by a pair of the sub graphs, denoted as *cut(*A_i, A_j*)*

$$cut(A_i, A_j) = \sum_{v_i \in A_i, v_j \in A_j} w_{ij}$$

— The size of each sub graph. The size of a graph A can be measured by the number of the nodes denoted as $|A|$ or by the total weights of its edges denoted as *vol(*A*)*.

Furthermore, in the cases that the number of the sub graphs equals 2 (K = 2), A_1, A_2 can be presented as *A* and $\bar{A}$ where $\bar{A}$ = V $-$ *A* denotes the complement of *A* in V. Then *cut(*$A, \bar{A}$*)* = *cut(*$\bar{A}, A$*)* is used to measure total weights of the edges escaping from A.

2.2.2 Graph cut

The goal of graph clustering is to cut the graph into sub graphs so that the edges connecting nodes in different sub graphs have the low weights and the edges within the sub graphs have high weights. Mathematically, in the case that G is cut into K sub graphs, the goal is to minimizes the quantity.

$$cut(A_1, A_2, ..., A_k) = \sum_{i=1}^{K} cut(A_i, \bar{A}_i)$$

This problem is therefore called *mincut* problem which can be solved efficiently (Stoer & Wagner, 1997). However, in many real cases, the *mincut* solution separates just one node from the rest of the graph. Therefore, we need a solution to keep the number of nodes of each sub graph "reasonably large". The possible approaches are to take into account the size of the sub graphs in the cost function above. Correspondingly to two ways to measure the size of a graph, there are two approaches, *RatioCut* (Hagen & Kahng, 1992) and *Ncut* (Shi & Malik, 2000).

$$RatioCut(A_1, A_2, ..., A_k) = \sum_{i=1}^{K} \frac{cut(A_i, \bar{A}_i)}{|A_i|}$$

$$Ncut(A_1, A_2, ..., A_k) = \sum_{i=1}^{K} \frac{cut(A_i, \bar{A}_i)}{vol(A_i)}$$

Unfortunately, solving these problems are NP-hard problems (Wagner & Wagner, 1993). However, spectral clustering is a way to solve relaxed versions of these problems (Von Luxburg, 2007; Malliaros & Vazirgiannis, 2013).

2.2.3 Spectral clustering method

To solve the relaxed versions of the optimization of the cut-based graph clustering, the spectral clustering is based on Laplacian matrices *L* that has the unnormalized version formed from the *weight matrix* and *degree matrix* of the graphs.

$$L = D - W$$

The spectrum (eigenvectors) of this matrix is used to obtain the final clusters (or sub graphs) (Malliaros & Vazirgiannis, 2013). In particular, the spectral clustering can be used to cluster a graph into K sub graphs by following the steps below.

1. Compute unnormalized Laplacian matrix L
2. Compute the eigenvectors (denoted **x**) and eigenvalues (denoted λ) of matrix L by solving $L\mathbf{x} = \lambda\mathbf{x}$. These eigenvectors then are ordered by eigenvalues.
3. Build the matrix M that has K columns which are the first K eigenvectors obtained in step 2, M = $(x_1, x_2, ..., x_K)$
4. Perform k-means algorithm to cluster the rows of the matrix M into K clusters which correspond to K sub graphs.

However, this algorithm does not take into account the sizes of the sub graphs (or *mincut* solution). Therefore, normalized Laplacian matrices, which can be L_{rw} or L_{sym}, are used to replace the unnormalized Laplacian matrix (Shi & Malik, 2000; Hamad & Biel, 2008; Ng *et al.*, 2001)

$$L_{rw} = D^{-1}L$$
$$L_{sym} = D^{-1/2}LD^{-1/2}$$

Moreover, as the last step of the spectral clustering method is to apply k-means algorithm which uses the distance to cluster the matrix M, it can also be helpful if matrix M is normalized before performing k-means (Ng *et al.*, 2001). One common normalization is row sums to have norm 1 as below

$$u_{ij} = \frac{v_{ij}}{\sum_k (v_{ik}^2)^{1/2}}$$

2.2.4 Modularity value and Louvain method

Another approach for density-based graph clustering is based on modularity. This modularity was actually designed to measure the strength of division of a graph into clusters (or communities). However, modularity is often used as the objective functions in graph clustering. A popular method of this approach is the Louvain method (Vincent *et al.*, 2008).

2.2.4.1 Modularity value

By definition, modularity is the fraction of the edges that fall within the given groups minus the expected fraction if edges were distributed at random. Given that an undirected graph *G* with the weight matrix W = (w_{ij}) is clustered into K sub graphs. The modularity of a sub graph A (denoted Q_A) will be :

$$Q_A = \frac{W_A}{W_G} - \frac{d_A * d_A}{2W_G * 2W_G}$$

in which, W_G, W_A are the total weights of graph G and sub graph A respectively and d_A is the total degrees of the nodes in A.

$$W_G = \frac{1}{2}\sum_{v_i, v_j \in G} w_{ij}$$

$$W_A = \frac{1}{2}\sum_{v_i, v_j \in A} w_{ij}$$

$$d_A = \sum_{v_i \in A} d_i = \sum_{v_i \in A}\sum_j w_{ij}$$

The strength of division of a graph into K sub graphs, or graph clustering modularity, is the total of modularities (now denoted as Q_i) of all the sub graphs

$$Q = \sum_{i=1}^{K} Q_i$$

This graph clustering modularity will have values ranging from -1 to 1 and higher modularity values indicate better graph clustering (Vincent *et al.*, 2008).

2.2.4.2 Louvain method

In the Louvain method, modularity is used as the objective function in graph clustering. In particular, the idea of the Louvain method is that the nodes will be moved around to their neighbor clusters so that the modularity of the graph clustering increases. This Louvain method can have several phases :

At the first phase, the method firstly considers each node of the graph as an individual cluster. That means at the beginning, the number of clusters equals the number of nodes of the graph. Each node v_i has a number of neighbors v_j if there is an edge $\{v_i, v_j\}$. The Louvain method works by moving every node v_i from its cluster to all the neighboring clusters, which contain at least one neighbor node v_j, for maximum gain of modularity. To illustrate how this step works, let Q_{ib} and Q_{ia} be the modularities of the cluster containing node v_i respectively before and after removing node v_i from it. Similarly, Q_{jb} and Q_{ja} are the modularities of the neighbor cluster containing the neighbor node v_j respectively before and after adding node v_i into that neighbor cluster. The gain of modularity (denoted ΔQ) by moving node v_i from its cluster to the neighbor cluster of node v_j is calculated by the following formula :

$$\Delta Q = (Q_{ia} + Q_{ja}) - (Q_{ib} + Q_{jb})$$

By calculating ΔQ with all the neighbor clusters, node v_i will be placed in the cluster that gives the maximum of ΔQ, which must also be a positive number. In case all the ΔQ are negative numbers, node v_i will stay in its original cluster. This first phase terminates when no movement of nodes can increase the modularity.

The second phase of the method starts by building a new graph using the result of the first phase. In particular, in the new graph, each cluster now is considered as a node (denoted cluster node). The weight of the edge connecting two cluster nodes equals the total weights of all the edges connecting two nodes in the two clusters. In addition, the nodes in the same clusters will create a self-loop edge whose weight is the total of the weights of all the edges connecting two nodes inside that cluster.

After building the new graph, the steps taken in the first phase are repeated to cluster the new graph. The question is how many phases we should take to cluster a graph ? The answer is that it depends on the needs. More specifically, if we want more clusters, we can stop at an earlier phase, first phase for instance. We also can run the algorithm until the new graph cannot be clustered.

One main issue we need to consider in the case of large graphs is the computation time. To improve the computation time, the Louvain method also presented some simple heuristics such as stopping the first phase when the gain of modularity is below a given threshold (Vincent *et al.*, 2008). Another approach for the heuristics is that instead of moving a node to all of its neighbor clusters, we move it to the neighbor clusters of a certain number of neighbor nodes after ordering them by weights of the connecting edges. For example, in our dataset (section 2.1), the number of the neighbor nodes we tried is 10 and it returns the same result as the one where no heuristics is applied.

2.2.4.3 Customization of Louvain method

In our work, we need to add several constraints to this method. One constraint is that each final hospital cluster must contain a public University Hospital (fr. Centre Hospitalier Universitaire - CHU). This constraint is taken into account in our implementation by customizing

the Louvain method. In particular, these CHUs are considered as "seed" nodes of the graph. Our customization method is that at the steps of moving nodes to their neighboring clusters for the maximum gain of modularity, these "seed" nodes will not be moved. Instead, the other nodes will be moved to neighbor clusters which contain the "seed" nodes.

2.3 Evaluation for hospital communities

The modularity value is the first criterion to be used to evaluate the graph clustering methods. The higher modularity values indicate the better results in graph clustering. On the other side, since we are grouping the hospitals into communities for the purpose of effectively sharing medical records, we also use the percentage of previous hospitals visited outside the communities to evaluate the efficiency of the methods. This percentage value indicates the rate at which hospitals would not have access to the patients' medical record from previous hospitalization if the communities were created. Therefore, the methods with smaller values for this percentage are better. There is also the fact that the number of hospitals in each community has an impact on the two values above. For example, a community structure that has one very big community containing almost all the hospitals while other communities containing only one hospital naturally gives the highest values for the percentage value above. Therefore, the balance in terms of number of hospitals in each communities should be taken into account when the evaluations are conducted.

3 Results and Discussions

3.1 Implementation approaches

In the literature, there are already libraries implementing both graph clustering methods introduced in the previous section. For example, *sklearn* library and *python-louvain* library have already implemented *SpectralClustering* and *Louvain* method respectively. However, as mentioned before, besides comparing the performances of the two methods, we need to customize them so that we can add the constraints to meet our needs. Therefore, we have implemented these methods using the programming language of Python 3 and the environments of Anaconda 3 and Ubuntu 18.

3.2 Method comparison

To measure the performance, we use three criteria mentioned in section 2.3. In particular, the table 2 below contains the values for the three criteria of the graph clustering methods. However, it should be noted that in table 2, the number of clusters is 19 which is generated by the Louvain method after running three phases. For comparison purposes, we use the same number of 19 for all the spectral clustering methods.

TABLE 2 – *Graph clustering method comparison.* $\boldsymbol{SC}$ *: Spectral clustering with unnormalized Laplacian matrix L.* $\boldsymbol{SC_{sym}}$ *: Spectral clustering with normalized Laplacian matrix* $\boldsymbol{L_{sym}}$*.* $\boldsymbol{SC_{rw}}$ *: Spectral clustering with normalized Laplacian matrix* $\boldsymbol{L_{rw}}$*.* $\boldsymbol{SC_{rw+norm}}$ *: Spectral clustering with normalized Laplacian matrix* $\boldsymbol{L_{rw}}$ *and the matrix M (of the first K eigenvectors) is also normalized.* $\boldsymbol{LV}$ *: Louvain method*

Evaluation criteria	SC	SC_{sym}	SC_{rw}	$SC_{rw+norm}$	LV
Modularity value	0.000	0.701	0.804	0.816	0.822
% previous hospitals outside community	0.006	20.44	9.33	10.32	9.84
# hospitals in biggest community	1,758	838	379	269	260
# hospitals in smallest community	1	2	2	22	22

As it can be seen in table 2, the SC method (or *mincut* solution) does not work on our dataset. In particular, it returns a very big community covering almost all hospitals (1,758

TABLE *3 – Details of communities by Louvain method.* ***C*** *: Community.* ***NoH*** *: Number of hospitals.* ***NoP*** *: Number of patient flows. % : of patient flows within community*

C	NoH	NoP	%
1	260	2,186,805	90.45
2	171	1,171,110	92.78
3	138	967,312	91.06
4	133	637,415	88.34
5	125	529,591	89.93
6	102	638,005	88.57
7	95	751,129	90.16
8	94	892,364	95.60
9	85	422,831	85.73
10	83	455,684	84.61
11	78	580,529	90.22
12	77	529,977	89.80
13	66	463,606	91.41
14	60	426,626	92.74
15	59	321,093	85.57
16	48	289,226	82.88
17	47	313,069	89.54

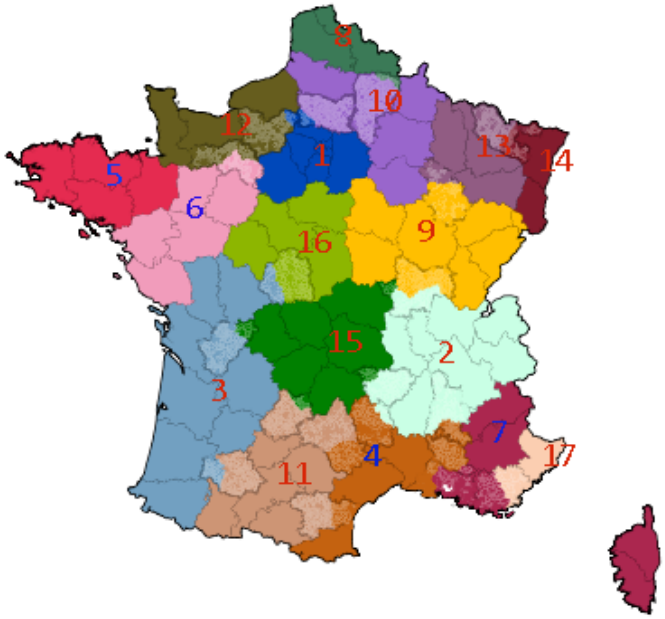


FIGURE *1 – Locations of communities. Lighter colors indicate "split" departments*

over 1,777) while in the other communities, the numbers of hospitals are just 1 or 2. In addition, between SC_{rw} and SC_{sym} methods, the SC_{rw} method returns better results in all the criteria. Moreover, by normalizing matrix M (mentioned in section 2.2.3), the corresponding method labeled $SC_{rw+norm}$ returns a higher value for the modularity as well as more balanced communities. Particularly, the modularity increases from 0.804 to 0.816 while the number of hospitals in the biggest community reduces from 379 to 269 and the number of hospitals in the smallest community increases from 2 to 22. The only criteria for which $SC_{rw+norm}$ method is not better than SC_{rw} is the percentage (%) of previous hospitals outside the community. These values are 10.32 and 9.33 for $SC_{rw+norm}$ and SC_{rw} respectively. This result can be explained by the number of patient flows inside the biggest community by each method. In particular, the $SC_{rw+norm}$ method returns the biggest community that has 269 hospitals and the numbers of patient flows inside this community is 2,248,178 (17.17%) while the numbers that SC_{rw} method returns are 379 and 2,892,368 (22.09%) respectively. Therefore, although compared with SC_{rw} method, $SC_{rw+norm}$ method returns a higher rate of hospitals that cannot access the patients' medical records from previous hospitalizations, we can still conclude that $SC_{rw+norm}$ method is better in this case. Finally, to help us select the better method, we compare $SC_{rw+norm}$ method with the Louvain method. As it can be seen in table 2, the Louvain method returns better results in all the criteria. Therefore, we have selected the Louvain method to cluster hospitals into communities which share medical records.

3.3 Final result

As mentioned in the previous section, we have selected the Louvain method for our work. After phase 3, the algorithm returns 19 communities among which the first biggest 17 communities are in metropolitan France. The details of these 17 communities are provided in Table 3. In addition to this table, a spatial map is used to visualize these 17 communities.

As the map (Figure 1) shows, the two biggest communities (community 1 and 2) in term of both the number of hospitals and the number of patient flows inside are located in Paris and Lyon, which are the biggest cities of France, and their nearby regions. In addition, 19 over 96 departments in metropolitan France are split into at least two different communities. The

"split" departments are shown in the map with lighter colors compared to the color presenting the communities. Moreover, 15 over 132 GHTs [3] are split into different communities. This knowledge can be used to advise health authorities that they should not use administrative region borders as constraints when creating hospital communities.

4 Conclusions

Lacking medical information from the previous hospitalizations about a patient can prevent hospitals from providing effective and high-quality treatments to those patients. Therefore, building hospital communities among which medical records are shared is needed. Based on the dataset of patient flows between hospitals, we approach graph clustering methods to effectively group the hospitals into communities. After comparing the performance with spectral clustering methods, we have selected the Louvain method for our work. As a result, after running three phases of the Louvain method, we obtained 19 hospital communities. Among them, the 17 biggest communities are in metropolitan France. More importantly, some departments in metropolitan France as well as some GHTs are split into at least two different communities. This knowledge confirms the limitations of building hospital communities based on administrative boundaries. In addition, such methods could be used to effectively design groups of hospitals that should share common electronic medical records.

Références

BOUDEMAGHE T. & BELHADJ I. (2017). Data resource profile : The french national uniform hospital discharge data set database (pmsi). *Int. J. Epidemiol.*, **46(2)**, 392–392d.

HAGEN L. & KAHNG A. (1992). New spectral methods for ratio cut partitioning and clustering. *IEEE Trans, Computer-Aided Design*, **11(9)**, 1074–1085.

HAMAD D. & BIEL P. (2008). Introduction to spectral clustering. In *ICTTA 2008*, p. 1–6.

HAREL D. & KOREN Y. (2001). On clustering using random walks. In *FST TCS 2001 : Foundations of Software Technology and Theoretical Computer Science.*, p. 18–41.

LAM W. W. M. & CHAN K. (2008). A graph mining algorithm for classifying chemical compounds. In *2008 IEEE International Conference on Bioinformatics and Biomedicine*, p. 321–324.

MALLIAROS F. & VAZIRGIANNIS M. (2013). Clustering and community detection in directed networks : A survey. *ArXiv*, **abs/1308.0971**.

NG A., JORDAN M. & WEISS Y. (2001). On spectral clustering : analysis and an algorithm. In *Advances in Neural Information Processing Systems 14*, p. 849–856.

SHI J. & MALIK J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22(8)**, 888–905.

SOUZA J., SILVA E., BRITO P., COSTA J., SALGADO A. & MEIRA S. (2013). Using graph clustering for community discovery in web-based social networks. In *ICSI 2013 : Advances in Swarm Intelligence*, p. 120–129.

STOER M. & WAGNER F. (1997). A simple min-cut algorithm. *J. ACM*, **44(4)**, 585–591.

VINCENT B. D., GUILLAUME J. L., LAMBIOTTE R. & LEFEBVRE E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech*, p. P10008.

VON LUXBURG U. (2007). A tutorial on spectral clustering. *Stat Comput*, **17**, 395–416.

WAGNER D. & WAGNER F. (1993). Between min cut and graph bisection. In *In Proceedings of the 18th International Symposium on Mathematical Foundations of Computer Science (MFCS)*, p. 744–750.

3. There are no patient flows inside 5 GHTs