Corpus annoté de cas cliniques en français

Natalia Grabar[1] Cyril Grouin[2] Thierry Hamon[2, 3] Vincent Claveau[4]
(1) CNRS, UMR 8163 ; Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000, Lille, France
(2) LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France
(3) Université Paris 13, Sorbonne Paris Cité, F-93430 Villetaneuse, France
(4) Univ Rennes, Inria, CNRS, IRISA, F-35000, Rennes, France
natalia.grabar@univ-lille.fr, cyril.grouin@limsi.fr,
thierry.hamon@limsi.fr, vincent.claveau@irisa.fr

RÉSUMÉ

Les corpus textuels sont utiles pour diverses applications de traitement automatique des langues (TAL) en fournissant les données nécessaires pour leur création, adaptation ou évaluation. Cependant, dans certains domaines comme le domaine médical, l'accès aux données est rendu compliqué, voire impossible, pour des raisons de confidentialité et d'éthique. Il existe néanmoins de réels besoins en corpus cliniques pour l'enseignement et la recherche. Pour répondre à ce défi, nous présentons dans cet article le corpus *CAS* contenant des cas cliniques de patients, réels ou fictifs, que nous avons compilés. Ces cas cliniques en français couvrent plusieurs spécialités médicales et focalisent donc sur différentes situations cliniques. Actuellement, le corpus contient 4 300 cas (environ 1,5M d'occurrences de mots). Il est accompagné d'informations (discussions des cas cliniques, mots-clés, etc.) et d'annotations que nous avons effectuées au regard des besoins de la recherche en TAL dans ce domaine. Nous présentons également les résultats de premières expériences de recherche et d'extraction d'information qui ont été effectuées avec ce corpus annoté. Ces expériences peuvent fournir une *baseline* à d'autres chercheurs souhaitant travailler avec les données.

ABSTRACT

Annotated corpus with clinical cases in French.

Textual corpora are important for several NLP tasks because they provide suitable information for designing, adapting and evaluating these NLP applications. Yet, in some domains, such as the medical one, for confidentiality and ethical reasons, access to representative data is complicated or even impossible. Still, real need exists for this kind of corpora, both for training and research. In this paper, we propose the *CAS* corpus in French containing clinical cases of patients, real or fake. They cover various medical specialities and focuse on different clinical situations. Currently, the corpus contains 3,600 cases (almost 1.3M word occurrences). This corpus is associated with additional information (discussions of clinical cases, key-words...) and annotations that we produced to answer common research issues in this domain. We also present results from preliminary experiments of information retrieval and extraction performed on this corpus. These experiments can provide a *baseline* for the researchers interested in working with these data.

MOTS-CLÉS : Corpus clinique, cas clinique, annotations, catégorisation, extraction d'information.

KEYWORDS: Clinical corpus, clinical case, annotations, categorization, information extraction.

1 Introduction

Les corpus textuels sont utiles pour diverses tâches et applications du traitement automatique des langues (TAL) car ils fournissent les informations nécessaires pour la création, l'adaptation et l'évaluation des applications et d'outils. Cependant, dans certains domaines, pour des raisons de confidentialité et d'éthique, l'accès aux données représentatives devient très compliqué voire impossible. Les domaines du médical et du juridique relèvent de cette situation : dans le domaine juridique, l'information sur les procès et les délibérations reste confidentielle, tandis que dans le domaine médical les dossiers cliniques de patients sont aussi confidentielles car le secret médical doit être respecté. Dans les deux cas, les données ne peuvent pas être utilisées en dehors du cadre initial, en raison de la présence de données nominatives.

Notons que depuis plusieurs années déjà, les outils et méthodes d'anonymisation et de désidentification sont devenus disponibles et fournissent des résultats compétitifs (Ruch *et al.*, 2000; Sibanda & Uzuner, 2006; Uzuner *et al.*, 2007; Grouin & Zweigenbaum, 2013) en atteignant jusqu'à 90 % de précision et de rappel. Ces outils ont été développés pour traiter des textes en plusieurs langues et provenant de différents domaines. Leur exploitation pourrait donc aider les chercheurs à accéder aux données sensibles. Cependant, les données désidentifiées peuvent aussi être difficiles à obtenir et à utiliser pour la recherche car il a été noté que le risque de ré-identification des personnes persiste. Cela concerne par exemple les patients (Meystre *et al.*, 2014; Grouin *et al.*, 2015) dont les histoires médicales peuvent être uniques. D'autres difficultés d'ordre institutionnel ou juridique peuvent également complexifier la situation et l'accès aux données. Pour ces diverses raisons, la désidentification des données personnelles n'est souvent pas suffisante pour pouvoir les exploiter dans les contextes de recherche et d'enseignement en dehors des structures hospitalo-universitaires.

Néanmoins, il existe de réels besoins en développement de méthodes et outils visant les applications orientées sur des domaines spécialisés. Il en est ainsi dans le domaine médical, où des outils de recherche et d'extraction d'information sont nécessaires, par exemple pour le recrutement et l'inclusion de patients dans des essais cliniques, la recherche de patients similaires, le codage PMSI, etc. De manière plus fondamentale, il s'agit de tâches comme par exemple l'indexation des dossiers cliniques, l'étude de la temporalité, de l'incertitude ou de la négation, l'extraction des traitements prescrits ou des effets indésirables, etc. (Embi *et al.*, 2005; Hamon & Grabar, 2010; Uzuner *et al.*, 2011; Fletcher *et al.*, 2012; Sun *et al.*, 2013; Campillo-Gimenez *et al.*, 2015; Kang *et al.*, 2017). Ces questions de recherche sont communément abordées en langue anglaise, qui dispose de corpus dédiés, mais restent fragiles dans d'autres langues, comme le français, faute de corpus disponibles et accessibles pour la recherche.

Un autre point crucial, qui motive grandement notre travail, concerne la fiabilité des outils et la reproductibilité des résultats avec des données similaires provenant de sources différentes ou même avec des données provenant du même type de sources. Les travaux de recherche du domaine biomédical souffrent ainsi d'une vive critique en raison du manque de reproductibilité des résultats obtenus (Chapman *et al.*, 2011; Collins & Tabak, 2014; Cohen *et al.*, 2016). Une première étape vers la reproductibilité passe par la disponibilité d'outils, de corpus et de données de référence.

Dans ce travail, nous nous focalisons sur la création d'un corpus disponible contenant des données issues ou proches des données cliniques. L'objectif de cet article consiste à présenter le corpus de cas cliniques en français, les annotations actuellement disponibles et quelques premières expériences et leurs résultats. Dans ce qui suit, nous présentons d'abord quelques travaux sur la création de corpus médicaux en mettant l'accent sur les corpus disponibles pour la recherche (section 2). Nous décrivons

ensuite le corpus de cas cliniques en français (section 3) que nous proposons, les annotations actuelles (section 4) et les expériences qu'il a permis d'effectuer jusqu'ici (sections 5 à 7). Nous concluons en indiquant quelques directions de travaux futurs (section 8).

2 Corpus cliniques disponibles librement

Dans le domaine médical, nous pouvons distinguer deux principaux types de corpus : scientifiques et cliniques. Les *corpus scientifiques* proviennent de la littérature scientifique. Ils deviennent de plus en plus disponibles pour la recherche grâce aux initiatives de publications ouvertes, comme celles soutenues par la NLM (National Library of Medicine) dans le portail PUBMED [1] spécifiquement dédié au domaine biomédical, ou des portails généralistes comme HAL [2] et ISTEX [3]. Certains corpus scientifiques fournissent des annotations et catégorisations précises. Ils sont souvent créés pour des compétitions TAL (Kelly *et al.*, 2013; Goeuriot *et al.*, 2014) ou proviennent des travaux de chercheurs (Tsuruoka *et al.*, 2005; Szarvas *et al.*, 2008).

En ce qui concerne les *corpus cliniques*, ils sont liés aux événements cliniques des patients (histoire médicale, soins médicaux, prescriptions, analyses de laboratoires, procédures chirurgicales, etc.). Il est compliqué d'avoir un accès libre à ce type de données pour des raisons évoquées plus haut (données nominales et sensibles, risque de ré-identification, contexte institutionnel...).

Le présent article s'intéresse à ce dernier type de corpus et notre revue de la littérature porte sur les corpus cliniques librement disponibles pour la recherche :

— Le corpus *MIMIC* (Medical Information Mart for Intensive Care), actuellement dans sa troisième version, fournit le plus grand ensemble de données cliniques, structurées et non structurées, en anglais. MIMIC III provient d'une seule institution et contient les informations relatives aux patients qui y sont admis. Ces données concernent les examens médicaux, médicaments prescrits, résultats de laboratoire, encodage des actes et des diagnostics, rapports d'imagerie, durée de séjour à l'hôpital, etc. Ce corpus est exploité dans de nombreuses applications académiques et industrielles, pour la recherche, pour l'amélioration des soins et pour l'enseignement (Johnson *et al.*, 2016), satisfaisant ainsi toute la palette des contextes propres au domaine biomédical. Plusieurs travaux de recherche utilisent ces données pour la prédiction de la mortalité (Anand *et al.*, 2018; Feng *et al.*, 2018), l'identification du diagnostic et le codage (Perotte *et al.*, 2014; Li *et al.*, 2018), l'étude de la temporalité (Che *et al.*, 2018) ou encore la recherche de cas similaires (Gabriel *et al.*, 2018). Les données de ce corpus ont notamment été utilisées dans plusieurs compétitions de TAL, dont nous décrivons plusieurs ici : I2B2, N2C2, CLEF-eHEALTH.

— *I2B2* (Informatics for Integrating Biology and the Bedside) [4] est une compétition dont l'objectif est de motiver le développement et l'évaluation d'outils du TAL sur les données cliniques. Les données exploitées sont en anglais et désidentifiées. Les différentes éditions ont proposé des annotations spécifiques sur la désidentification, l'identification de fumeurs, les informations liées aux médicaments, les relations sémantiques entre entités, ou la temporalité (Uzuner, 2008; Uzuner *et al.*, 2011; Sun *et al.*, 2013).

1. https://www.ncbi.nlm.nih.gov/pubmed
2. https://hal.archives-ouvertes.fr/
3. https://www.istex.fr/
4. https://www.i2b2.org/NLP/DataSets/Main.php

— *N2C2* (National NLP Clinical Challenges) [5] a lieu depuis 2018. La compétition porte par exemple sur l'inclusion de patients dans les essais cliniques, la détection des effets indésirables provoqués par le prise de médicaments, la détection de similarités textuelles, l'extraction de l'histoire familiale de maladies ou la normalisation de concepts . Cette compétition a pris le relais de I2B2, tout en proposant des tâches plus complexes et plus ancrées dans la réalité et l'activité clinique.
— *CLEF-eHEALTH* [6] a connu plusieurs éditions : la détection de maladies et la normalisation des abréviations en 2013 et 2014, le traitement des notes d'infirmiers australiens en 2016, l'extraction des causes de décès dans les certificats de décès en français issus du CépiDc [7] en 2016 et 2017.
— Le défi *eHealth-KD* 2019 [8] vise à modéliser la langue utilisée dans les documents cliniques en espagnol et à traiter automatiquement ces documents. Deux tâches sont proposées : identifier et classer des séquences clés, puis détecter les relations sémantiques entre séquences.

Finalement, les données médicales proches des données cliniques peuvent aussi être trouvées dans les protocoles d'essais cliniques. Des exemples de ce type de corpus comportent les annotations d'informations sur les valeurs numériques en anglais (Claveau *et al.*, 2017), et de négation en français et portugais brésilien (Dalloux *et al.*, 2018).

3 Corpus de cas cliniques en français

Nous proposons le corpus nommé *CAS* qui contient des cas cliniques rédigés en français. Les cas cliniques décrivent les situations cliniques de patients, réels désidentifiés ou fictifs. Ils sont publiés dans différentes sources de données (scientifique, didactique, associatif, juridique...). Ils sont anonymisés au moment de la publication. Les cas cliniques ont pour objectif de présenter des situations cliniques typiques, dans un objectif didactique, ou des situations rares et complexes (propriété rencontrée dans des cadres scientifique et juridique). La figure 1 présente un exemple de cas clinique. Nous observons que les informations fournies sont de nature diverse : genre et âge du patient, motif de la consultation ou de l'hospitalisation, observation et résultats d'examens cliniques, résultats d'examens biologiques, traitements effectués (traitements chirurgicaux dans l'exemple de la figure), évolution de la maladie. En ceci, le contenu des cas cliniques est vraiment très proche du contenu des dossiers cliniques et en offre donc un bon exemple.

Une première version du corpus a été présentée dans une publication antérieure (Grabar *et al.*, 2018). Depuis, le corpus a été fondamentalement enrichi. Actuellement, le corpus global contient pas loin de 4 300 cas, soit presque 1 500 000 occurrences de mots. Le contenu provient de différentes sources (littérature scientifique, matériel didactique, support des associations, affaires juridiques) et représente différentes spécialités médicales (cardiologie, urologie, oncologie, obstétrique, pneumologie, gastro-entérologie, gériatrie, pharmacologie, etc.). En fonction de la spécialité, l'accent est mis sur des aspects différents (diagnostic d'une maladie, prise en charge, intervention chirurgicale, interactions médicamenteuses, etc.) et les cas peuvent aussi bien relater toute l'histoire de la maladie des patients que de se focaliser sur un épisode donné. Les cas recensés ont été publiés dans différents pays francophones (France, Belgique, Suisse, Canada, pays africains, pays tropicaux, etc.). Il s'agit donc

5. https://n2c2.dbmi.hms.harvard.edu/
6. https://sites.google.com/site/shareclefehealth/
7. http://www.cepidc.inserm.fr/
8. https://knowledge-learning.github.io/ehealthkd-2019

B.A., âgé de 36 ans, sans antécédents notables, a été admis en février 1994 pour des douleurs lombaires droites évoluant dans un contexte d'altération de l'état général. L'examen clinique avait montré une tension artérielle à 10/06 mm Hg chez ce patient apyrétique, avec un examen abdominal et neurologique normal par ailleurs. Les examens biologiques montraient un taux de globules blancs à 7000/mm3 , une créatinine à 8 mg/l et une glycémie à 0,90 g/l. L'abdomen sans préparation ne montrait pas de calcifications et l'échographie abdominale avait montré une masse latéro-vertébrale droite refoulant le rein droit vers l'extérieur (Figure 1). La tomodensitométrie abdominale (Figures 2 et 3) avait objectivé une formation tissulaire isodense arrondie de 5 cm de diamètre située en plein parenchyme du muscle psoas droit. Une biopsie échoguidée de la tumeur n'avait pas ramené de tissu tumoral. L'intervention menée par une lombotomie avait découvert une tumeur encapsulée, bien limitée de 5 cm de grand diamètre incluse dans le muscle psoas. Une tumorectomie complète était réalisée. A la coupe, la tumeur présentait un aspect blanchâtre fasciculé, de consistance ferme. A l'examen microscopique, on avait trouvé une prolifération de cellules fibroblastiques fusiformes sans anomalies cytologiques agencées en faisceaux dissociés par l'oedème et du tissu conjonctif comportant des petits capillaires, concluant à un fibrome. L'évolution a été bonne avec un recul de 4 ans.

FIGURE 1 – Un exemple de cas clinique (les références à des figures font partie du document)

de productions effectuées en français et décrivant des situations cliniques assez typiques et réelles de patients susceptibles de venir en consultation ou en hospitalisation dans un hôpital francophone. Les cas cliniques sont écrits par les médecins : les mêmes personnes qui écrivent les dossiers hospitaliers des patients.

Par ailleurs, les cas peuvent bénéficier de différents types d'annotations, comme présenté dans la section 4, et être associés avec d'autres types d'information. Les informations associées dépendent des sources d'où proviennent les cas. Par exemple, les cas cliniques publiés dans la littérature scientifique sont souvent accompagnés d'une discussion et des mots-clés, les cas cliniques provenant du matériel didactique peuvent être accompagnés de questions de contrôle des connaissances, alors que les cas provenant des affaires juridiques sont typiquement associés avec les jugements et pénalités.

4 Annotations du corpus

Les annotations que nous présentons concernent un sous-ensemble du corpus composé de 717 cas cliniques (soit 232 000 occurrences de mots). Comme l'ensemble du corpus, ces cas couvrent différentes spécialités médicales et proviennent de plusieurs sources. Ce corpus de 717 cas cliniques a été mis à disposition de la compétition DEFT 2019 [9]. En dehors des annotations présentées plus bas (les informations démographiques (section 4.1) et les informations cliniques générales (section 4.2)), les cas sont annotés manuellement avec des informations sémantiques plus fines (maladies, signes et symptômes, médicaments, procédures, dates, examens cliniques et biologiques, etc.) et annotés automatiquement avec des informations linguistiques (étiquetage morpho-syntaxique). Nous faisons également le bilan quantitatif des annotations démographiques et cliniques générales (section 4.3).

9. https://deft.limsi.fr/2019/

4.1 Informations démographiques

Les informations démographiques couvrent l'*âge* et le *genre* des patients. Les portions textuelles permettant d'en déterminer les valeurs sont annotées : la valeur numérique et l'unité pour les âges *(2 mois et demi, 36 ans, la quarantaine)*, les valeurs réelles *(sexe féminin, garçon)* ou les indices linguistiques permettant de les inférer : participes passés *(hospitalisé, intubée)*, pronoms personnels ou démonstratifs, déclencheurs *(M., Mme)*, expressions *(le patient, cette patiente)*. Les valeurs obtenues sont normalisées sous la forme d'un entier pour l'âge et des valeurs "féminin" ou "masculin" pour le genre (il n'existe aucun cas d'hermaphrodisme ou de dysgénésie).

4.2 Informations cliniques générales

Les informations cliniques générales concernent l'*origine* de la consultation (pathologie, signe ou symptôme qui se trouvent à l'origine de la consultation ou de l'hospitalisation décrites dans le cas) et l'état du patient à l'*issue* de l'hospitalisation (guérison, amélioration, stabilité, détérioration, décès). Lorsque le cas clinique intègre l'histoire de la maladie avec plusieurs épisodes d'hospitalisations ou de consultations, c'est le dernier épisode qui est retenu comme origine de la consultation ou de l'hospitalisation décrite et par rapport à laquelle une issue peut être définie.

4.3 Statistiques

Classe	*Annotateur 1/Annotateur 2*	*Annotateur 1/consensus*	*Annotateur 2/consensus*
âge	0,9844	0,9887	0,9944
genre	0,8044	0,9903	0,8143
issue	0,4654	0,6204	0,8152
origine	0,8734	0,8886	0,9755

TABLE 1 – Accords inter-annotateurs (F-mesure) calculés avec BRATeval : comparaison des portions pour *âge* et *origine*, et des valeurs normalisées pour *genre* et *issue*.

Le corpus a été annoté par deux annotateurs de manière indépendante. Le tableau 1 fournit les accords inter-annotateurs (F-mesure) calculés avec l'outil BRATeval. L'évaluation porte sur la portion annotée pour les classes *âge* et *origine*, et sur les valeurs normalisées pour les classes *genre* (valeurs possibles : masculin, féminin) et *issue* (valeurs possibles : guérison, amélioration, stable, détérioration, décès). Un consensus a permis de corriger les erreurs et oublis d'annotations. En cas de désaccords sur les frontières, qui concernaient la classe *origine*, la portion la plus englobante est conservée. Les désaccords sur l'issue concernent des valeurs proches (guérison/amélioration, stable/amélioration), des oublis d'annotation, ou des absences volontaires d'annotation dues à la difficulté de choisir la bonne valeur. Nous observons que la classe *issue* est moins simple qu'il n'y paraît, suscitant de nombreuses discussions lors du consensus. Globalement, nous pouvons voir que : (1) l'accord entre les deux annotateurs est proche de la perfection pour l'âge (0,9844), (2) l'accord est très bon pour le genre et l'origine (0,8044 et 0,8734), et (3) l'accord est faible pour l'issue (0,4654). Pour cette dernière catégorie, les valeurs de l'accord des deux annotateurs par rapport au consensus indiquent que chacun des annotateurs a fait des erreurs ou omissions d'annotation, de même que des annotations correctes qui ont été retenues dans la version consensuelle de l'annotation.

Classe	*Valeurs normalisées et nombre d'occurrences*
âge	0-9 ans (56), 10-19 ans (63), 20-29 ans (100), 30-39 ans (109), 40-49 ans (99), 50-59 ans (132), 60-69 ans (75), 70-79 ans (54), 80-89 ans (14), 90-99 ans (4), âge inconnu (21)
genre	féminin (321), masculin (418)
issue	guérison (227), amélioration (256), stabilité (55), détérioration (23), décès (117)

TABLE 2 – Statistiques d'annotations du corpus de cas cliniques

Le tableau 2 donne la répartition des valeurs normalisées des classes *âge*, *genre* et *issue* dans le corpus. Certaines catégories (80-89 ans et 90-99 ans pour l'âge, et détérioration pour l'issue) sont sous-représentées par rapport à d'autres.

Grâce à ses annotations et ses informations, le corpus CAS permet de tester des applications utiles dans le domaine clinique, comme la catégorisation et l'extraction d'information. Nous avons défini trois tâches sur la base des annotations produites et des informations disponibles dans le corpus :

1. association des mots-clés avec les cas cliniques (section 5),
2. association des cas cliniques et des discussions (section 6),
3. extraction d'information clinique (section 7).

Pour chacune de ces tâches, nous proposons des techniques simples, se voulant des systèmes permettant de fournir des résultats initiaux (*baseline*). Ces systèmes se veulent simples dans leur conception (techniques bien connues) et dans la mesure où ils n'utilisent pas de connaissances externes. Elles sont présentées dans les sections suivantes.

5 Association des mots-clés avec les cas cliniques

La première tâche consiste à associer des mots-clés à chacun des cas cliniques. Avec le développement des systèmes d'information hospitalière, la recherche de dossiers cliniques devient un réel défi pour les praticiens qui désirent trouver un dossier particulier ou un patient donné dans la masse des informations existantes dans un hôpital. L'indexation des dossiers cliniques s'impose alors comme une étape préalable à la recherche d'information.

L'entrée de cette tâche est un ensemble de cas cliniques avec leurs discussions, un ensemble des mots-clés possibles et le nombre de mots-clés attendus. En effet, dans le contexte clinique, l'indexation ou le codage de dossiers médicaux sont souvent effectués de manière contrôlée en exploitant des terminologies médicales existantes. Dans notre tâche, la vérité-terrain, ou les données de référence, est constituée avec les mots-clés assignés par les auteurs eux-mêmes aux publications scientifiques qu'ils ont écrites et d'où proviennent les cas cliniques. Un mot-clé peut être associé à plusieurs cas cliniques, un cas cliniques peut recevoir un à plusieurs mots-clés, et certains mots-clés de l'ensemble ne doivent pas être associés aux cas cliniques ou leurs discussions.

Nous avons 290 cas cliniques/discussions dans le jeu d'entraînement et 213 cas cliniques/discussions dans le jeu de test. Dans les deux cas, l'ensemble de mots-clés regroupe les mots-clés de l'entraînement et du test et contient 1 311 mots-clés.

5.1 Évaluation

Nous avons constitué des données d'entraînement (290 cas avec leurs mots-clés), pour permettre l'emploi de méthodes d'apprentissage, et de test (213 cas). L'évaluation de cette tâche prend en compte la possibilité de produire une liste ordonnée de mots-clés candidats, du plus pertinent au moins pertinent. Pour confronter cette liste ordonnée à la liste de référence, nous utilisons deux mesures classiquement utilisées en Recherche d'Information : la moyenne des R-Précisions (précision mesurée au rang N, $Pr(N)$, où N est le nombre de mots-clés attendus pour ce cas clinique) et la MAP (*Mean Average Precision*, moyenne de l'*Average Precision*; voir formule 1).

$$MAP = \frac{1}{|C|} \sum_{c \in \mathcal{C}} \frac{\sum_{k=1}^{m} Pr(k) * \mathbb{1}_{t(k) \in Rel(c)}}{N} \quad (1)$$

5.2 Systèmes de référence

A titre de comparaison, nous avons produit deux systèmes de référence, ou des *baselines*. Le tableau 3 indique les performances de ces systèmes de référence.

Système	*MAP*	*R-précision*
Baseline 1	0,177	0,236
Baseline 2	0,434	0,428

TABLE 3 – Résultats des deux *baselines* pour la tâche 1 d'association de mots-clés aux cas cliniques et leurs discussions.

Un premier système de référence consiste à rechercher les mots-clés de la liste à l'identique dans chaque couple cas clinique/discussion, puis à sélectionner les mots-clés dont la fréquence d'utilisation dans le couple cas clinique/discussion est la plus élevée. En cas de fréquences identiques entre plusieurs mots-clés (par exemple, une fréquence de 1), nous conservons les mots-clés les plus longs, en émettant l'hypothèse qu'un mot-clé long est plus significatif qu'un mot-clé court. Enfin, nous limitons le nombre de mots-clés retournés au nombre de mots-clés attendu. Notons que cette approche donne la possibilité d'exploiter les cas cliniques et/ou la discussion. Les résultats indiquent que le traitement séparé du cas et de la discussion, avec une sélection des mots-clés a posteriori, est plus efficace que la fusion des deux. Cette *baseline* obtient une MAP de 0,177 et une R-précision de 0,236 sur les données de test.

Le deuxième système de référence exploite la pondération Okapi-BM25 (Robertson *et al.*, 1998) pour ordonner les candidats mots-clés. Cette pondération permet ainsi de tenir compte de la fréquence des mots-clés dans le cas clinique traité, mais aussi dans l'ensemble des cas cliniques. Tous les mots-clés fournis sont cherchés dans les documents et ceux identifiés sont pondérés par BM-25. La liste retournée correspond ainsi aux mots-clés trouvés, ordonnée par le score BM-25 décroissant. Cette approche obtient une MAP de 0,434 et une R-précision de 0,428 sur les données de test.

6 Association des cas cliniques et des discussions

La deuxième tâche consiste à associer la discussion au cas clinique correspondant, ce qui peut se révéler utile pour les médecins qui veulent identifier dans la littérature scientifique des observations cliniques similaires à celles de leurs patients. Une telle recherche bibliographique vise à trouver les méthodes de diagnostic ou de traitement les plus appropriées. L'entrée de cette tâche est un ensemble de cas cliniques et un ensemble de discussions. Chaque cas doit être associé à une discussion. Une discussion donnée peut être associée à plus d'un cas clinique.

Nous avons 290 cas (et leurs discussions) dans le jeu d'entraînement et 213 cas (et leurs discussions) dans le jeu de test. Il existe donc des doublons au sein des discussions.

6.1 Évaluation

Pour cette tâche, une seule réponse est attendue : une seule discussion à associer à un cas clinique. En revanche, comme indiqué plus haut, une même discussion peut concerner plusieurs cas. L'évaluation se fait classiquement par les mesures du rappel et de la précision, calculées globalement (c'est-à-dire, formellement, macro-précision et macro-rappel) : si le système renvoie une réponse pour tous les cas, ces deux mesures sont égales. Le script d'évaluation gère les doublons qui se trouvent au sein des discussions : il suffit qu'une discussion, parmi les discussions doublons correctes, soit associée à un cas clinique.

6.2 Systèmes de référence

Système	*Précision*	*Rappel*
Baseline	0,9500	0,9500

TABLE 4 – Résultats de la *baseline* pour la tâche 2 d'association des cas cliniques et des discussions.

L'approche *baseline* que nous proposons consiste à calculer les similarités entre toutes les discussions et tous les cas cliniques. Ces derniers sont simplement représentés comme des sacs-de-mots. La similarité utilisée est de nouveau Okapi-BM25. On obtient ainsi une matrice de similarité entre cas et discussions. A ce point, une discussion peut être plus proche du cas c_1 que que du cas c_2, en terme du score BM-25, et être la première classée pour c_2 et la cinquième pour c_1. L'attribution optimale se fait alors en utilisant l'algorithme hongrois (Kuhn & Yaw, 1955) sur cette matrice de similarités. La précision (équivalente au rappel) obtenue par cette approche, sur le jeu de test, est de 0,95 (tableau 4).

7 Extraction d'information clinique

Les annotations manuelles disponibles sur ce corpus fournissent la possibilité d'effectuer d'autres expériences qui sont également proches des besoins cliniques en traitement d'informations. Il s'agit typiquement de l'extraction d'information pour la recherche de patients avec un profil donné ou par rapport aux critères d'inclusion dans les protocoles d'essais cliniques. Pour cette expérience, quatre informations sont annotées et recherchées : l'âge, le genre, l'issue et la portion de texte expliquant

la raison d'admission du patient. L'entrée de cette tâche est un ensemble de cas cliniques. Un cas clinique concerne en général un patient mais certains cas peuvent être concernés par plus d'un patient. Il n'est pas nécessaire d'associer les informations extraites (par exemple, l'âge et le genre) entre elles.

Le jeu d'entraînement contient 290 cas cliniques et le jeu de test 427 cas cliniques. Chaque cas est en général annoté avec les quatre types d'information.

7.1 Évaluation

Pour cette tâche, les quatre types d'informations à extraire sont évalués selon deux protocoles, en fonction de la nature de l'information :

— L'âge, le genre, et l'issue sont classiquement évalués par la précision et le rappel.

— L'admission est représentée par une portion de texte ou, dans quelques cas, par plusieurs portions de texte. Pour comparer la portion attendue à celle(s) prédite(s), nous utilisons plusieurs mesures. Nous calculons les valeurs de rappel et précision sur les mots de ces portions de textes. Ces mesures peuvent être faites au niveau de chaque cas et moyennées, ou calculées globalement sur l'ensemble des cas, résultats en micro-précision et micro-rappel, ou macro-précision et macro-rappel. Nous proposons également de mesurer l'intersection en nombre de mots entre la portion attendue et la portion prédite, normalisée par l'ensemble des mots de la référence et de la prédiction. Cette mesure effectuée pour chaque cas est ensuite moyennée, résultant ainsi dans la mesure que nous appelons *micro-overlap*, définie formellement dans la formule 2.

$$micro - overlap = \frac{1}{|C|} \sum_{c \in \mathcal{C}} \frac{\sum_{k=1}^{m} \frac{TP}{TP + FP + FN}}{N} \tag{2}$$

7.2 Systèmes de référence

Nous avons produit deux systèmes de référence. Les résultats obtenus par ces deux systèmes se trouvent dans le tableau 5.

	Système à base de règles		Apprentissage supervisé	
Classe	*Précision*	*Rappel*	*Précision*	*Rappel*
âge	0,7897	0,7685	0,9608	0,9116
genre	0,9138	0,9014	0,9602	0,9535
issue	0,4444	0,4247	0,5321	0,5246
origine (micro)	0,4182	0.0061	0,7707	0,5559
origine (macro)	0,0321	0.0163	0,5141	0,5647

TABLE 5 – Résultats des méthodes de baseline pour la tâche 3 : extraction de l'âge, du genre, de l'issue, et de l'origine d'admission

Un premier système de référence repose sur un ensemble limité de règles propres à chaque catégorie : 5 règles pour le genre, 9 pour l'âge, et 7 pour l'issue. Il exploite pour ceci une liste de termes

(*femme, homme, madame, monsieur...*), les parties anatomiques genrées, les pronoms personnels pour compléter l'identification du genre. Sur la catégorie *Origine*, ce système est limité à seulement six règles pour traiter les portions commençant par la préposition "pour" suivi de termes ressemblant à des signes ou symptômes (*pour des épisodes fugaces de palpitations*, *pour une gêne respiratoire*, etc.). Ce travail limité et rapide ne peut donner cependant lieu à des résultats viables sur cette dernière catégorie. Comme indiqué dans le tableau 5, ce système a des performances élevées pour le genre et l'âge. En revanche les performances des deux autres catégories restent basses, surtout en ce qui concerne le rappel.

Un deuxième système de référence proposé exploite des approches par apprentissage artificiel. Deux approches sont exploitées :

— Le genre et l'issue sont considérés comme des problèmes de catégorisation de texte. Nous utilisons un algorithme de Régression Logistique dans lequel nous représentons le texte sous la forme de sac-de-mots, avec une simple pondération TF *(term frequency)*. Le modèle est appris sur le jeu d'entraînement et utilisé ensuite pour prédire le genre et l'issue pour les cas du jeu de test.

— Pour l'âge et l'admission, il s'agit de repérer dans les documents les portions faisant mention de l'âge et de la raison d'admission. Ils ont donc été considérés comme des problèmes d'étiquetage. Les textes sont étiquetés en parties-du-discours et lemmatisés avec TagEx [10]. Pour l'entraînement, les informations d'âge et d'admission sont projetés sur le document sous la forme d'étiquette IOB. Nous entraînons ensuite un modèle CRF (implémenté par Wapiti (Lavergne *et al.*, 2010)) sur ces données qui est ensuite appliqué au données du jeu de test.

Les résultats obtenus sont indiqués dans le tableau 5. Nous voyons que deux catégories (âge et genre) montrent des performances très élevées, étant supérieures à 0,90 en termes de précision et de rappel. Les deux autres catégories ont des performances un peu plus modestes mais qui restent élevées : entre 0,50 et 0,77. Pour ce système aussi, le rappel est plus difficile à gérer que la précision.

8 Conclusion

Nous avons décrit un corpus de cas cliniques en français, qui correspondent à des données proches de celles créées et utilisées dans le contexte hospitalier. Actuellement, le corpus contient 4 300 cas cliniques (environ 1,5M d'occurrences de mots). Une partie du corpus (717 cas cliniques) a été annotée avec quatre types d'informations sémantiques (âge et genre du patient, origine de consultation et issue de consultation). L'accord inter-annotateurs, calculé avec la F-mesure, est supérieur à 0,80 pour trois catégories (âge, genre et origine) et est de 0,4654 pour la catégorie issue. Cette dernière a en effet présenté des difficultés d'annotation. De plus, ces 717 cas cliniques sont également associés avec les mots-clés et une discussion, tous les deux étant fournis par les publications d'origine.

Le corpus de cas cliniques de patients décrit dans cet article peut être utilisé pour l'enseignement et la recherche. Ainsi, plusieurs cadres d'évaluation de tâches de catégorisation et d'extraction d'information sont en cours de développement, montrant ainsi le potentiel que CAS représente pour la recherche. La mise à disposition de ce corpus pour la compétition DEFT 2019 [11] en fait partie,

10. TagEx est un outil d'étiquetage morpho-syntaxique et de lemmatisation développé à l'IRISA et disponible en web-service : `https://allgo.inria.fr`. Il est adapté au traitement de documents du domaine biomédical.

11. (`https://deft.limsi.fr/2019/`

alors que les résultats des systèmes de référence ont vocation de fournir des données de comparaison par rapport auxquelles d'autres systèmes pourront se positionner.

De manière plus générale, nous pensons que la disponibilité de ce corpus et des annotations vont stimuler la recherche sur les données de type clinique en langue française. Ceci va contribuer à garantir la reproductibilité des résultats et la robustesse des méthodes et outils.

Nous prévoyons d'enrichir le corpus avec d'autres cas cliniques et de fournir d'autres annotations consensuelles. Ce corpus et ses annotations pourront donc faire objet d'autres compétitions TAL.

Remerciements

Ce travail a bénéficié d'une aide de l'État attribuée au labex COMIN LABS et gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'avenir » portant la référence ANR-10-LABX-07-01. Ce travail s'inscrit également dans le projet *CLEAR (Communication, Literacy, Education, Accessibility, Readability)* financé par l'ANR sous la référence ANR-17-CE19-0016-01. Nous remercions les relecteurs pour leurs remarques constructives.

Références

ANAND R., STEY P., JAIN S., BIRON D., BHATT H., MONTEIRO K., FELLER E., ML R., IN S. & ES C. (2018). Predicting mortality in diabetic icu patients using machine learning and severity indices. In *AMIA Jt Summits Transl Sci Proc*, p. 310–319.

CAMPILLO-GIMENEZ B., BUSCAIL C., ZEKRI O., LAGUERRE B., LE PRISÉ E., DE CREVOISIER R. & CUGGIA M. (2015). Improving the pre-screening of eligible patients in order to increase enrollment in cancer clinical trials. *Trials*, **16**(1), 1–15.

CHAPMAN W. W., NADKARNI P. M., HIRSCHMAN L., D'AVOLIO L. W., SAVOVA G. K. & UZUNER O. (2011). Overcoming barriers to nlp for clinical text : the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc*, **18**(5), 540–543.

CHE Z., PURUSHOTHAM S., CHO K., SONTAG D. & LIU Y. (2018). Recurrent neural networks for multivariate time series with missing values. *Sci Rep*, **8**(1), 6085.

CLAVEAU V., SILVA OLIVEIRA L. E., BOUZILLÉ G., CUGGIA M., CABRAL MORO C. M. & GRABAR N. (2017). Numerical eligibility criteria in clinical protocols : annotation, automatic detection and interpretation. In *AIME (Artifical Intelligence in Medicine in Europe)*.

COHEN K. B., XIA J., ROEDER C. & HUNTER L. E. (2016). Reproducibility in natural language processing : A case study of two r libraries for mining pubmed/medline. In *LREC Int Conf Lang Resour Eval*, p. 6–12.

COLLINS F. & TABAK L. (2014). Nih plans to enhance reproducibility. *Nature*, **505**, 612–613.

DALLOUX C., CLAVEAU V., GRABAR N. & MORO C. (2018). Portée de la négation : détection par apprentissage supervisé en français et portugais brésilien. In *TALN 2018*, p. 1–6.

EMBI P., JAIN A., CLARK J. & HARRIS C. (2005). Development of an electronic health record-based clinical trial alert system to enhance recruitment at the point of care. In *Ann Symp Am Med Inform Assoc (AMIA)*, p. 231–35.

FENG M., MCSPARRON J., KIEN D., STONE D., ROBERTS D., SCHWARTZSTEIN R., VIEILLARD-BARON A. & CELI L. (2018). Transthoracic echocardiography and mortality in sepsis : analysis of the mimic-iii database. *Intensive Care Med*, **44**(6), 884–892.

FLETCHER B., GHEORGHE A., MOORE D., WILSON S. & DAMERY S. (2012). Improving the recruitment activity of clinicians in randomised controlled trials : A systematic review. *BMJ Open*, **2**(1), 1–14.

GABRIEL R., KUO T., MCAULEY J. & HSU C. (2018). Identifying and characterizing highly similar notes in big clinical note datasets. *J Biomed Inform*, **82**, 63–69.

GOEURIOT L., KELLY L., LI W., PALOTTI J., PECINA P., ZUCCON G., HANBURY A., JONES G. & MÜLLER H. (2014). Share/clef ehealth evaluation lab 2014, task 3 : User-centred health information retrieval. In *CLEF*, Lecture Notes in Computer Science (LNCS), p. 43–61 : Springer.

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). Cas : French corpus with clinical cases. In *LOUHI 2018*, p. 1–12, Bruxelles, Belgique.

GROUIN C., GRIFFON N. & NÉVÉOL A. (2015). Is it possible to recover personal health information from an automatically de-identified corpus of French EHRs ? In *Proc of LOUHI*, Lisbon, Portugal.

GROUIN C. & ZWEIGENBAUM P. (2013). Automatic de-identification of french clinical records : Comparison of rule-based and machine-learning approches. In *Stud Health Technol Inform, Proc of MedInfo*, volume 192, p. 476–80, Copenhagen, Denmark.

HAMON T. & GRABAR N. (2010). Linguistic approach for identification of medication names and related information in clinical narratives. *J Am Med Inform Assoc*, **17**(5), 549–54.

JOHNSON A. E., POLLARD T. J., SHEN L., WEI H. LEHMAN L., FENG M., GHASSEMI M., MOODY B., SZOLOVITS P., CELI L. A. & MARK R. G. (2016). MIMIC-iii, a freely accessible critical care database. *Scientific Data*, **3**(160035), 1–9.

KANG T., ZHANG S., TANG Y., HRUBY G. W., RUSANOV A., ELHADAD N. & WENG C. (2017). EliIE : An open-source information extraction system for clinical trial eligibility criteria. *J Am Med Inform Assoc*, **24**(6), 1062–1071.

KELLY L., GOEURIOT L., SUOMINEN H., MOWERY D. L., VELUPILLAI S., CHAPMAN W. W., ZUCCON G. & PALOTTI J. (2013). Overview of the share/clef ehealth evaluation lab 2013. In *CLEF*, Lecture Notes in Computer Science (LNCS) : Springer.

KUHN H. W. & YAW B. (1955). The hungarian method for the assignment problem. *Naval Research Logistic Quarterly*, **2**, 83–97.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 504–513 : Association for Computational Linguistics.

LI M., FEI Z., ZENG M., WU F., LI Y., PAN Y. & WANG J. (2018). Automated ICD-9 coding via a deep learning approach. In *IEEE/ACM Trans Comput Biol Bioinform*.

MEYSTRE S., SHEN S., HOFMANN D. & GUNDLAPALLI A. (2014). Can physicians recognize their own patients in de-identified notes ? In *Stud Health Technol Inform 205*, p. 778–82.

PEROTTE A., PIVOVAROV R., NATARAJAN K., WEISKOPF N., WOOD F. & ELHADAD N. (2014). Diagnosis code assignment : models and evaluation metrics. *J Am Med Inform Assoc*, **21**, 231–237.

ROBERTSON S. E., WALKER S. & HANCOCK-BEAULIEU M. (1998). Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive. In *Proc. of the 7th Text Retrieval Conference, TREC-7*, p. 199–210.

RUCH P., BAUD R. H., RASSINOUX A.-M., BOUILLON P. & ROBERT G. (2000). Medical document anonymization with a semantic lexicon. In *Ann Symp Am Med Inform Assoc (AMIA)*, p. 729–733, Los Angeles, CA.

SIBANDA T. & UZUNER O. (2006). Role of local context in de-identification of ungrammatical, fragmented test. In *NAACL-HLT 2006*, New York, USA.

SUN W., RUMSHISKY A. & UZUNER Ö. (2013). Evaluating temporal relations in clinical text : 2012 i2b2 challenge. *JAMIA*, **20**(5), 806–813.

SZARVAS G., VINCZE V., FARKAS R. & CSIRIK J. (2008). The BioScope corpus : annotation for negation, uncertainty and their scope in biomedical texts. In *BIONLP*, p. 38–45.

TSURUOKA Y., TATEISHI Y., KIM J., OHTA T., MCNAUGHT J., ANANIADOU S. & TSUJII J. (2005). Developing a robust part-of-speech tagger for biomedical text. *LNCS*, **3746**, 382–392.

UZUNER O. (2008). Second i2b2 workshop on natural language processing challenges for clinical records. In *Ann Symp Am Med Inform Assoc (AMIA)*, p. 1252–3.

UZUNER O., LUO Y. & SZOLOVITS P. (2007). Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc*, **14**, 550–563.

UZUNER O., SOUTH B. R., SHEN S. & DUVALL S. L. (2011). 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*, **18**(5), 552–556.