Segmentation automatique des rapports médicaux en utilisant les réseaux de neurones convolutionnels

Walid Zeghdaoui*,** Frederik Joly**
Omar Boussaid* Fadila Bentayeb*

*Université de Lyon, Université Lyon 2, ERIC EA 3083
5, Av. Pierre Mendès-France, 69676 Bron, France
{walid.zeghdaoui, omar.boussaid, fadila.bentayeb}@univ-lyon2.fr
**Sword Group - 9, Av. Charles de Gaulle, 69370 St-Didier-au-Mont-d'Or, France
{walid.zeghdaoui, frederik.joly}@sword-group.com

Résumé. L'un des défis majeurs de la médecine de précision est d'orienter la recherche et le développement de solutions thérapeutiques spécifiques en exploitant les informations contenues dans les rapports médicaux. Ces rapports sont souvent maintenus sous forme de textes non structurés et constituent un volume important de données. L'extraction de connaissances à l'aide des techniques de traitement automatique de langage naturel (TALN) pourrait donc aider à améliorer les soins de santé et la prise de décisions médicales. Cette étape est souvent précédée par une phase de segmentation de textes pour bien identifier les parties du rapport d'un grand intérêt. Dans cet article, nous présentons notre système automatique de segmentation de rapports médicaux en sections sémantiques et qui consiste en deux étapes complémentaires. D'abord un algorithme basé sur la détection de titres pour l'identification de certaines sections. La deuxième partie est une tâche de classification de phrases supervisée basée sur des algorithmes d'apprentissage profond. Ce système a été testé sur 500 rapports et a atteint une précision de classification de plus de 96%.

1 Introduction

En dépit des efforts consacrés pour la saisie de données cliniques dans un format structuré, la plupart des rapports médicaux sont aujourd'hui maintenus en textes libres. La quantité de ces données a augmenté considérablement au cours de la dernière décennie et risque de se multiplier dans les années à venir. Ces données contiennent une mine d'informations et représentent une riche base de connaissances qui doit être impérativement exploitée pour améliorer le processus de soins de santé. En effet, l'utilisation de ces informations permet d'aider les médecins sur plusieurs plans, notamment faciliter la prise de décisions médicales. Par exemple l'identification du diagnostic le plus approprié pour un patient, mais aussi pour faire progresser la recherche clinique, car la communauté médicale s'efforce constamment de trouver de nouveaux moyens pour mener des recherches qui s'inscrivent dans la lutte contre les maladies. Afin de faciliter leur tâche, notre travail consiste à rendre ces informations facilement accessibles. Ainsi l'utilisation des techniques de recherche et d'extraction d'information est devenue

inévitable pour l'automatisation de plusieurs tâches. En effet, ces données sont volumineuses et les traitements manuels deviennent de plus en plus fastidieux voire même impossibles dans certains cas d'usage. Cependant, le format non structuré de ces rapports rend difficile l'extraction d'informations significatives, d'autant plus lorsqu'ils sont longs et que seules quelques parties sont intéressantes pour certains utilisateurs. Par exemple, les médecins pourraient s'intéresser à des informations spécifiques comme les antécédents chirurgicaux d'un patient ou encore la conclusion d'un médecin dans un rapport, tandis que d'autres utilisateurs pourraient se rapporter aux données personnelles (noms, adresses, ...etc.) à des fins d'anonymisation.

La segmentation des rapports médicaux en sections joue donc un rôle clé non seulement parce qu'elle nous permet de les diviser en unités significatives, de façon à ce qu'on puisse en extraire une partie spécifique en réponse à une requête donnée d'un utilisateur, mais aussi parce qu'elle nous aide à mieux comprendre le sens des informations qui y sont stockées. Par exemple, les codes *SNOMED CT*[1] sont souvent représentés sur 5 chiffres, de même que les codes postaux en France. Donc l'utilisation seule des techniques classiques de reconnaissance d'entités nommées ne suffit pas puisque ces codes peuvent avoir des significations différentes. Ce problème est aussi connu dans la littérature sous le nom de désambiguïsation d'entités nommées. Dans les rapports médicaux, les codes postaux sont généralement utilisés au début dans une section dédiée aux informations personnelles, contrairement aux codes *SNOMED CT* qui sont souvent utilisés dans les conclusions en fin de rapport, il nous est donc facile de distinguer entre ces termes et de surmonter ce genre de problèmes à l'aide d'un système de segmentation.

Dans le cadre de ce travail, nous traitons les rapports médicaux de quatre institutions membres des 20 Centres de Lutte Contre le Cancer (CLCC). Il s'agit respectivement de l'Institut Paoli-Calmettes (IPC) à Marseille, de l'Institut Curie (IC) à Paris, de l'Institut du Cancer de Montpellier (ICM) et du Centre Léon-Bérard (CLB) à Lyon. Par ailleurs, nous avons également l'intention de déployer notre système de segmentation dans d'autres centres. Par conséquent, l'identification de types de rapports devient compliquée puisqu'il n'existe actuellement aucun modèle universel pour les rapports médicaux écrits en France.

Compte tenu de ces questions, nous proposons dans cet article un système scalable de segmentation automatique des rapports médicaux en sections prédéfinies. Nous rapportons ensuite les résultats de performance sur 500 rapports tirés aléatoirement de quatre institutions membres des Centres de Lutte Contre le Cancer.

2 Travaux connexes

La segmentation automatique de textes (i.e. le processus qui consiste à diviser un texte en unités cohérentes) est un problème fondamental dans le traitement du langage naturel, la classification des documents et la recherche d'informations, et a fait l'objet de plusieurs travaux de recherche pour différents types de langages et d'applications allant de la génération automatique de résumés de documents à l'application des politiques de sécurité. Les systèmes de segmentation efficaces fonctionnent généralement sur du texte dans des domaines et des sources spécifiques. Par exemple, le traitement du texte utilisé dans les rapports médicaux est un problème très différent du traitement des articles de presse ou des annonces immobilières.

1. *SNOMED CT* est la terminologie clinique la plus complète et la plus précise au niveau international, permettant aux professionnels de la santé et aux chercheurs d'adopter un langage commun.

Koshorek et al. (2018) ont abordé le problème de segmentation de texte comme une tâche d'apprentissage supervisé, où chaque phrase est dénotée par une étiquette marquant la fin ou non d'une section. Ils ont présenté un modèle neuronal composé de deux réseaux de neurones récurrents (RNN) basés sur l'architecture LSTM (Hochreiter et Schmidhuber, 1997). Dans leur article, ils ont aussi introduit un nouvel ensemble de données, WIKI-727K, dédié à l'entraînement des modèles de segmentation de texte.

En revanche, Glavaš et al. (2016) ont présenté une approche non supervisée pour la segmentation de texte en construisant un graphe dont les nœuds représentent des phrases, et les arêtes relient les phrases sémantiquement similaires. Les sections cohérentes sont ensuite déterminées en trouvant les cliques maximales du graphe de similarité.

Diverses méthodes de segmentation de texte ont émergé au cours des dernières années, cependant, moins d'efforts ont été consacrés à son application dans le domaine clinique.

Apostolova et al. (2018) ont développé un système de segmentation basé sur un classificateur SVM (Cortes et VAPNIK, 2009) qui divise un texte en huit unités sémantiques. Leur modèle a été entraîné sur des rapports de radiologie qui sont des notes de patients externes et bénéficient d'une structure très concise contrairement à d'autres types de rapports cliniques.

Ganesan et Subotin (2015) ont proposé un modèle supervisé utilisant la régression logistique avec une approche de combinaison de contraintes capable de reconnaître l'en-tête, le pied de page et toutes les sections de niveau supérieur d'un texte clinique. Cette méthode fonctionne au niveau de la ligne plutôt qu'au niveau de la phrase, ce qui pourrait générer une séquence d'étiquettes qui n'a pas de sens.

Parmi le peu de travaux qui se sont focalisés sur les rapports médicaux écrits en langue française, Deléger et Névéol (2014) ont présenté une approche qui consiste à séparer le contenu médical des documents et les informations administratives contenues dans les en-têtes et pieds de page, en entraînant un modèle statistique à champs conditionnels aléatoires (CRF (Lafferty et al., 2002)).

Bien que les approches susmentionnées aient atteint une bonne précision, elles reposent sur la nature de certains types de rapports. Dans cet article, nous proposons un système scalable pour segmenter automatiquement les rapports médicaux en sections sémantiques.

3 Données et définition du problème

3.1 Définition des sections

Dans le cadre de nos travaux, nous avons tiré aléatoirement 500 rapports médicaux de 417 dossiers patients différents parmi un corpus de 191738 rapports issus de quatre institutions membres des CLCC. Nous avons sélectionné un échantillon de 100 rapports pour l'analyse préliminaire qui consiste à étudier le contenu des rapports et définir les différentes sections de manière efficace, nous permettant ainsi de bien exploiter ce découpage. Nous avons ainsi, et avec l'aide d'un expert, identifié 7 sections couvrant tous les rapports utilisés. Le Tableau 1 ci-dessous regroupe les noms des sections, leur description et le nombre total de phrases dans chaque section des 500 rapports.

Les antécédents personnels et familiaux regroupent plusieurs informations comme les interventions chirurgicales antérieures, les hospitalisations et les maladies chroniques chez les membres de la famille. Toute information relative à une personne physique susceptible d'être

Section	Description	Nombre
Antécédent personnel	Antécédents médicaux personnels	461 (10.9%)
Antécédent familial	Antécédents médicaux familiaux	206 (4.8%)
Donnée personnelle	Informations permettant d'identifier une personne	814 (19.2%)
Donnée bruyante	Informations à faible valeur médicale	524 (12.4%)
Recommandation	Recommandations et suivi médical	264 (6.2%)
Conclusion	Conclusion du médecin	167 (4%)
Contenu	Tout ce qui n'est pas dans les sections précédentes	1802 (42.5%)
Total	-	4238 (100%)

TAB. 1 – *Description et nombre des sections identifiées.*

identifiée, directement ou indirectement, par exemple les noms et les adresses, est considéré comme une donnée personnelle. Les données à faible valeur médicale sont les informations qui intéressent peu les médecins lors de la prise de décision médicale. Enfin, les recommandations sont les conseils et/ou les propositions des médecins aux patients, par exemple : «Prochain RDV dans 6 mois».

3.2 Challenges de la segmentation des rapports cliniques

Étant le texte libre le format couramment utilisé dans les rapports médicaux, la segmentation exige une compréhension des particularités de ces rapports. L'analyse préliminaire nous a montré qu'il est d'usage que les médecins utilisent des titres pour débuter des phrases ou des paragraphes. Nous avons donc pensé que l'identification de ces mots clés pourrait nous aider dans notre tâche de segmentation. Nous avons listé manuellement tous les titres contenus dans les 100 rapports sélectionnés pour l'analyse préliminaire et nous avons constaté que plus de 95% de ces titres ont des caractéristiques communes. Par exemple : ils ont une longueur maximale de 6 mots, commencent par un retour de ligne suivi d'une majuscule et se terminent par deux points et/ou un retour de ligne. La Figure 1 suivante permet de donner une idée.

Antécédent personnel : Antécédent de carcinome endométrioïde.

Résultats du bilan :
Confrontation au précédent scanner du 12 décembre 2012.

CONCLUSION
L'ensemble du bilan actuel n'a retrouvé aucune lésion évolutive. 80903

FIG. 1 – *Exemples de titres (**en gras**) dans les rapports.*

L'intuition d'utiliser des règles pour détecter les titres et ensuite étiqueter toutes les phrases regroupées dans le paragraphe suivant peut sembler évidente, mais l'utilisation seule des expressions régulières est loin d'être suffisante pour plusieurs raisons, par exemple :

1. Les phrases appartenant à des sections différentes peuvent être regroupées dans un même paragraphe, auquel cas, cette méthode nous induira en erreur.
2. De la même manière, on peut trouver des phrases d'une seule section réparties sur plusieurs paragraphes. Cela peut également nous induire à un manque de précision.
3. Les titres sont sujets à des fautes d'orthographe dues à des erreurs humaines et pourraient donc ne pas être détectés.
4. Et enfin, la plupart des rapports ne contiennent pas ou peu de titres explicites.

La Figure 2 ci-après illustre certaines des difficultés que nous avons listées.

Fait le 15/06/2004.
Cs Dr DUPONT

Résultast :
La cytoponction est en faveur d'une cytostéatonécrose.

A fait une échographie EV : pas d'anomalie particulière.

FIG. 2 – *Exemple des difficultés rencontrées avec les titres dans les rapports.*

Dans cet exemple, la première phrase du rapport ci-dessus n'est pas précédée par un titre, appartient à la section *Donnée bruyante* et est suivie d'une phrase qui appartient à une autre section, *Donnée personnelle*. La troisième phrase représente un titre qui fait référence à la section *Contenu*. Cependant, les expressions régulières ne sont pas suffisantes pour faire correspondre des chaînes mal orthographiées, dans ce cas précis, ("*Résultats*" au lieu de "*Résultast*"). Enfin, les deux dernières phrases appartiennent à la même section et sont réparties sur deux paragraphes différents (i.e. séparés par une suite d'au moins deux retours à la ligne consécutifs). L'utilisation des règles permet alors de détecter seulement la section de la première phrase.

Afin de surmonter les limitations d'un système de segmentation basé essentiellement sur des expressions régulières, nous avons opté à une solution hybride permettant ainsi de combiner une approche basée sur les règles, que nous allons définir dans la section suivante, avec une approche de classification de textes appliquée au niveau des phrases. Dans l'état de l'art actuel, les modèles de classification de textes sont basés sur l'apprentissage automatique supervisé (Mironczuk et Protasiewicz, 2018), cela leur permet de bien s'adapter et de se généraliser aux variations des phrases, et de faciliter la maintenabilité de tels systèmes.

3.3 Annotation des phrases

L'annotation des phrases consiste à attribuer pour chaque phrase un label ou une étiquette significative en fonction de son contenu. Dans le cadre de cette étude, nous avons utilisé un outil interne ayant une interface graphique qui permet d'annoter manuellement toutes les phrases de notre ensemble de données. Les étiquettes attribuées correspondent aux sections identifiées lors de l'analyse préliminaire. La qualité de l'annotation est très importante pour la construction d'un modèle de classification de phrases pleinement généralisable. De la même manière, les phrases annotées doivent être représentatives du corpus global.

4 Méthodes

Nous avons utilisé une approche hybride combinant à la fois un algorithme basé sur la détection des titres et un modèle supervisé de classification de phrases.

4.1 Algorithme basé sur des règles

À ce niveau, on effectue une première passe dans chaque rapport afin d'extraire des informations fiables et utiles pour l'étape suivante. Cette première étape est exclusivement basée sur la détection des titres. Chaque titre détecté est utilisé comme une règle pour attribuer une des sept sections que nous avons définies aux phrases du paragraphe qui suit le titre. Cependant, les règles de détection de titres ont une forte probabilité de produire de faux positifs (i.e. inclure des phrases qui satisfont toutes les contraintes liées à la définition d'un titre, et qui ne sont pas des titres). Afin de pallier à ce problème, nous avons procédé en deux temps :

D'abord on effectue une présélection de titres, c'est ce qu'on appellera par la suite des titres potentiels ou candidats, et qui doivent satisfaire toutes les contraintes suivantes :

1. Pas de titre vide.
2. Un titre contient au plus 6 mots (séparés par des espaces).
3. Un titre doit se tenir sur une seule ligne (i.e. ne contient pas de retour à la ligne).
4. Un titre doit commencer par une majuscule.
5. Un titre doit se terminer par un deux-points et/ou un saut de ligne.

Ensuite, on sélectionne les vrais titres en s'appuyant sur un dictionnaire construit lors de l'analyse préliminaire, contenant les expressions fréquemment utilisées dans les titres, et les sections auxquelles elles appartiennent. Le Tableau 2 illustre des exemples du dictionnaire.

Section	Expressions
Antécédent personnel	"ATCD", "antécédents personnels", "historique".
Antécédent familial	"ATCD Familiaux", "antécédent familial".
Donnée personnelle	"Prénom, "Adresse, "Numéro de téléphone".
Donnée bruyante	"Hôpital de Lyon", "Page 1/2".
Recommandation	"Prochain RDV dans 2 mois".
Conclusion	"conclusion", "au final".
Contenu	"Traitement", "Indication".

TAB. 2 – *Exemples d'expressions pour chacune des 7 sections.*

Par ailleurs, il faut tenir compte des variantes possibles qui n'ont pas encore été identifiées au cours de l'analyse préliminaire. Pour ce faire, nous avons introduit un processus de normalisation appliqué à tous les titres candidats :

1. Supprimer les caractères spéciaux (tiret, astérisque, etc.) ;
2. Supprimer les stopwords («de», «du», «la», etc.) ;
3. Transformer les suites de caractères vides en un seul espace ;
4. Remplacer les caractères accentués (par exemple, «é» par «e») ;

5. Mettre tous les titres candidats en minuscules.

Certaines sections ont des caractéristiques communes, y compris le format des titres utilisés. En effet, il est d'usage que les médecins utilisent un seul titre pour deux sections différentes. Par exemple, le titre "**Antécédent :**" peut faire référence à deux sections (*Antécédent personnel* et *Antécédent familial*. De la même manière, les médecins peuvent regrouper plusieurs phrases de section différentes dans un même paragraphe, par exemple, *Donnée personnelle* et *Donnée bruyante*. Nous avons donc décidé de regrouper temporairement certaines sections comme le montre le Tableau (3) ci-dessous.

Sections temporaires	Sections finales
Données	Donnée personnelle
	Donnée bruyante
Antécédents	Antécédent personnel
	Antécédent familial
Contenu	Contenu
Conclusion	Conclusion

TAB. 3 – *Regroupement de sections finales dans des sections temporaires.*

Ce choix a été motivé principalement par deux raisons :

1. Réduire le risque d'une mauvaise affectation d'une section à un titre.
2. Réduire le nombre de règles à maintenir.

Enfin, l'analyse préliminaire a également révélé que la majorité des conclusions ont des titres. L'identification de ces sections devient alors une tâche relativement facile et dépend de la détection de ces titres. Nous avons donc choisi de gérer les conclusions en utilisant seulement les règles, d'autant plus que les phrases qu'on y trouve n'ont pas de caractéristiques qui pourraient les distinguer de celles de la section *Contenu* par exemple.

4.2 Classification de textes par apprentissage automatique

La deuxième étape de notre système de segmentation a été modélisée comme une tâche de classification de phrases où on attribue à chaque phrase dans un rapport une des six sections parmi celles précédemment prédéfinies (les conclusions étant détectées lors de la première phase). Pour cela, nous avons mis en place un mécanisme naïf pour la détection des limites de phrases dans les rapports. Chaque rapport est donc découpé en plusieurs phrases, qui sont à leur tour normalisées avec le processus suivant :

1. Supprimer les caractères spéciaux (tiret, astérisque, etc.) ;
2. Remplacer les caractères accentués (par exemple, «é» par «e») ;
3. Transformer les suites de caractères vides en un seul espace ;
4. Transformer les phrases en minuscules.

Nous avons utilisé les phrases extraites de l'ensemble des comptes rendus pour tester et comparer les performances des meilleurs algorithmes d'apprentissage automatique de l'état de l'art dédiés à la classification de textes. Nous avons donc pu tester les arbres de décision et les machines à vecteurs de support avec *scikit-learn* [2], les réseaux de neurones peu profonds avec *fastText* [3] et les réseaux de neurones convolutionnels en s'appuyant sur *spaCy* [4] Chacune de ces méthodes tente de trouver lors de la phase d'entraînement le modèle qui fait correspondre le mieux entre les phrases et les labels (les sections), et qui permet de bien se généraliser aux phrases inconnues (i.e. les phrases qui n'appartiennent pas à l'ensemble d'entraînement).

4.2.1 Évaluation

Afin de réaliser nos expérimentations et entraîner les différents modèles en s'appuyant sur les meilleurs algorithmes d'apprentissage automatique dédiés à la classification de textes, nous avons constitué deux corpus de travail à partir des 500 rapports aléatoirement sélectionnés : un corpus d'entraînement de 200 rapports (50 rapports de chacun des 4 centres) et un corpus de test contenant les 300 rapports restants. Nous avons fait en sorte que les données d'entraînement et les données de test proviennent de la même distribution. Afin de permettre l'évaluation de la qualité de notre modèle, l'ensemble de test doit être suffisamment grand pour asseoir la crédibilité de la performance globale du système, d'autant plus lorsqu'on est confronté à un problème de données déséquilibrées où les classes (sections) sont représentées de manière complètement inéquitable.

La précision d'un modèle mesure sa capacité à classer **que** les documents pertinents (Exactitude). Le rappel d'un modèle mesure sa capacité à classer **tous** les documents pertinents (Exhaustivité). Ce sont deux mesures qui déterminent la pertinence d'un algorithme de classification. Nous avons évalué les résultats de nos différents modèles de classification entraînés en utilisant la précision (1), le rappel (2) et la F-mesure (3, avec $\beta = 1$).

$$Precision = \frac{Nombre\ de\ documents\ pertinents\ selectionnes}{Nombre\ total\ de\ documents\ pertinents} \tag{1}$$

$$Rappel = \frac{Nombre\ de\ documents\ pertinents\ selectionnes}{Nombre\ total\ de\ documents\ selectionnes} \tag{2}$$

$$F - mesure = \frac{(1+\beta^2) \times Precision \times Rappel}{\beta^2 \times Precision + Rappel} \tag{3}$$

Chaque algorithme a été utilisé pour entraîner plusieurs modèles en variant les hyperparamètres d'entrée à chaque exécution, pour choisir le modèle qui minimise le mieux la fonction de coût prédéfinie. Nous avons ainsi pu comparer les meilleurs modèles générés. Les performances de classification sont indiquées dans le Tableau (4) ci-dessous.

En prenant *F-mesure* comme critère d'évaluation, nous avons choisi de conserver *spaCy* pour cette tâche de classification de phrases puisqu'il permet d'obtenir les meilleures performances sur les phrases de nos comptes rendus.

2. *scikit-learn* : librairie dédiée à l'apprentissage automatique
3. *fastText* : librairie dédiée à l'apprentissage efficace de la représentation des mots et à la classification de phrases.
4. *spaCy* : librairie dédiée au traitement avancé du langage naturel.

Algorithme	Librairie	Précision	Rappel	F-mesure
DecisionTree	*scikit-learn*	73%	76%	72%
LinearSVCl2 (SVM)	*scikit-learn*	89%	77%	81%
FastText	*fastText*	84%	84%	84%
SpaCy	*spaCy*	**93**%	**91**%	**92**%

TAB. 4 – *Performances des différents algorithmes de classification.*

La classification de textes dans *spaCy* est basée sur les réseaux de neurones convolutionnels (CNN pour *Convolutional Neural Networks*), qui sont à la base dédiés au traitement d'images et spécialisés dans les tâches de reconnaissances de forme (Lecun et al., 1998). Ces architectures appliquées à des tâches de classification permettent souvent d'obtenir des résultats satisfaisants quand on dispose de suffisamment de données d'entraînement. Cela est dû à leur robustesse aux faibles variations d'entrée et au faible taux de prétraitement nécessaire à leur fonctionnement. L'architecture du CNN utilisée dans *spaCy* consiste en une succession de couches de convolution et d'agrégation (*Max-Pooling*) permettant l'extraction de caractéristiques, suivie d'une couche complètement connectée (*Fully Connected*) pour la classification.

La figure suivante permet de donner une intuition sur le fonctionnement des réseaux de neurones convolutionnels.

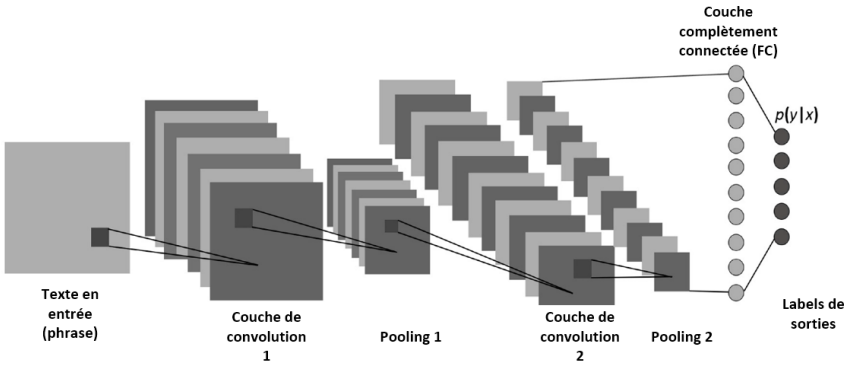

FIG. 3 – *Architecture classique des réseaux de neurones convolutionnels.*

Afin d'utiliser ces modèles sur du texte, *spaCy* se charge du *Word Embedding* (i.e. représenter chaque mot du dictionnaire par un vecteur de nombres réels correspondant dans l'espace vectoriel où sont définis tous les vecteurs) pour représenter dans l'ordre les mots de chaque phrase et ainsi pouvoir définir les paramètres de chaque couche de convolution : Le nombre de cartes de convolution, la taille des noyaux de convolution, et enfin, le schéma de connexion à la couche précédente. Chaque carte de convolution est le résultat d'une somme de convolution des cartes de la couche précédente par son noyau de convolution respectif. Un biais est

ensuite ajouté avant de passer le résultat à une fonction d'activation non-linéaire. Les couches d'agrégation permettent ensuite la réduction de la taille des cartes de telle sorte que les cartes d'activation de la dernière couche soient de taille 1. Cette dernière couche contient autant de neurones que de labels (sections) souhaités.

5 Expérimentations et résultats

Afin d'évaluer notre système de segmentation, nous avons combiné finalement l'algorithme basé sur les règles avec le modèle de classification *spaCy*, comme le montre la figure suivante.

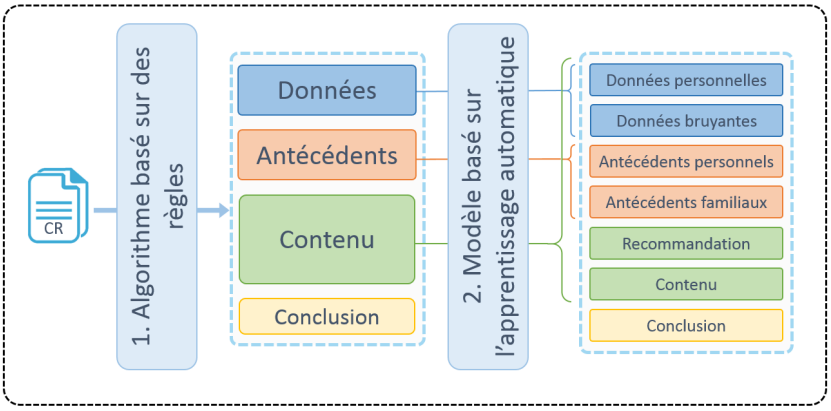

FIG. 4 – *Système de segmentation.*

Le modèle de classification *spaCy* prédit pour chaque phrase la suite des labels éventuels triés par probabilité décroissante. Nous avons utilisé le modèle de différentes manières en fonction de la section temporaire assignée aux phrases lors de la première étape de notre système. Par exemple, appliqué à la phrase «Nom : Nom de famille», le modèle pourrait générer la séquence de labels suivante : *(1) Contenu, (2) Donnée personnelle, (3) Donnée bruyante, (4) Antécédent familial, (5) Conclusion, (6) Antécédent personnel*. Dans ce cas, nous ne nous intéresserons qu'aux labels *Donnée personnelle* et *Donnée bruyante*. Puisqu'il s'agit de la section temporaire *Données*. De cette façon, le label le plus approprié est donc (*Donnée personnelle*). Pour les sections temporaires *Données* et *Antécédents*, seuls les labels de *Donnée personnelle*, *Donnée bruyante* et de *Antécédent personnel*, *Antécédent familial* respectivement sont concernés. Les résultats obtenus sont présentés dans le Tableau 5 suivant.

Section	Nombre	Exactitude
Antécédent personnel	254	91,81 %
Antécédent familial	133	92,8%
Donnée personnelle	279	96,81%
Donnée bruyante	271	96,38%
Recommandation	141	97,1%
Conclusion	79	99,2%
Contenu	979	97,2 %
Total	2136	96,06 %

TAB. 5 – *Performances du système de segmentation sur 300 rapports cliniques.*

Le taux de précision varie de 91,81% pour la section *Antécédent personnel* à 99,2% pour les sections *Conclusion* qui sont traitées uniquement lors de la première étape de notre système. Ces résultats montrent que la combinaison des techniques de traitement du langage naturel et des méthodes de classification des textes par apprentissage automatique pourrait être appliquée avec succès à la résolution de la segmentation automatique des rapports médicaux en texte libre. Les expériences révèlent également certains défis à relever. En fait, même après avoir effectué un réglage hyperparamétrique, «hyperparameter tuning» en anglais, pour chaque modèle utilisé dans notre tâche de classification de textes, il s'est avéré qu'un problème de biais élevé (une erreur due à des hypothèses erronées dans l'algorithme d'apprentissage) demeure. Cela confirme l'importance des règles que nous avons établies et la façon avec laquelle nous avons combiné les deux étapes de notre système.

Par ailleurs, il nous a été difficile de comparer les performances de notre système de segmentation avec les différentes approches identifiées dans les travaux connexes de manière générale puisque les tâches de segmentation de textes sont basées à la fois sur la modélisation informatique mais aussi l'étude linguistique, or la plupart des travaux qui traitent le même sujet, contrairement à notre approche, sont prédestinés à des rapports médicaux écrits en anglais. De plus, la première partie de notre système reste exclusivement adaptée aux types d'informations fréquemment rencontrées dans nos comptes rendus.

6 Conclusion et perspectives

La segmentation en sections des rapports médicaux en texte libre fournit des informations contextuelles importantes pour d'autres tâches d'extraction d'information automatisée. Cela pourrait améliorer le processus de soins de santé et faire progresser la recherche clinique. Dans cet article, nous avons proposé un système de segmentation automatique des rapports médicaux en deux temps. Tout d'abord, un algorithme basé sur des règles est développé pour identifier certaines sections en utilisant notre approche de détection des titres en deux phases. Cette première étape joue un rôle élémentaire de raffineemnt dans notre système de segmentation. En effet, l'algorithme fait une première passe sur les rapports pour extraire le maximum d'informations en tenant compte à la fois de la forme et du contenu de ces rapports. Ces informations

sont utilisées lors de la deuxième étape à l'exception des conclusions qui sont toutes identifiées en fin de la première étape. La deuxième partie consiste à former un modèle de classification de phrases par apprentissage automatique à l'aide de *spaCy* qui assigne une section à chaque phrase du rapport. Ce modèle est utilisé de manières différentes en fonction des résultats de la première étape. Le système a été testé sur 500 rapports annotés manuellement de quatre institutions membres des Centres de Lutte Contre le Cancer, et a réalisé une bonne performance. Ce système de segmentation est utilisé pour faciliter la recherche d'informations et l'extraction de connaissances à partir de rapports cliniques.

Il y a encore place à l'amélioration, notamment pour assurer la robustesse de notre système, puisqu'il est basé sur des caractéristiques définies lors de l'analyse préliminaire d'un échantillon aléatoire tiré d'un corpus plus important. Pour l'avenir, nous prévoyons de déployer notre système dans toutes les institutions des CLCC. Nous allons explorer également d'autres types de réseaux de neurones profonds comme les réseaux de neurones récurrents (Elman, 1990), en particulier les LSTM. En parallèle, nous explorons les méthodes d'annotation semi-automatique pour alléger le travail fastidieux d'annotation.

Références

Apostolova, E., D. Channin, D. Demner-Fushman, J. Furst, S. Lytinen, et D. Raicu (2018). Automatic segmentation of clinical texts-preliminary results.

Cortes, C. et V. VAPNIK (2009). Support-vector networks. *297*, 273–297.

Deléger, L. et A. Névéol (2014). Automatic identification of document sections for designing a french clinical corpus (identification automatique de zones dans des documents pour la constitution d'un corpus médical en français) [in french]. In *Proceedings of TALN 2014 (Volume 2 : Short Papers)*, pp. 568–573. Association pour le Traitement Automatique des Langues.

Elman, J. (1990). Finding structure in time. *14*, 179–211.

Ganesan, K. et M. Subotin (2015). A general supervised approach to segmentation of clinical texts. pp. 33–40.
English

Glavaš, G., F. Nanni, et S. P. Ponzetto (2016). Unsupervised text segmentation using semantic relatedness graphs. In **SEM 2016 : The Fifth Joint Conference on Lexical and Computational Semantics : proceedings of the conference ; August 11-12 2016, Berlin, Germany*, Stroudsburg, Pa., pp. 125–130. Association for Computational Linguistics.

Hochreiter, S. et J. Schmidhuber (1997). Long short-term memory. *9*, 1735–80.

Koshorek, O., A. Cohen, N. Mor, M. Rotman, et J. Berant (2018). Text segmentation as a supervised learning task.

Lafferty, J., A. Mccallum, et F. Pereira (2002). Conditional random fields : Probabilistic models for segmenting and labeling sequence data.

Lecun, Y., L. Bottou, Y. Bengio, et P. Haffner (1998). Gradient-based learning applied to document recognition. *86*, 2278 – 2324.

Mironczuk, M. et J. Protasiewicz (2018). A recent overview of the state-of-the-art elements of text classification.

Summary

One of the major challenges in precision medicine is to guide the research and development of specific therapeutic solutions through the extraction of knowledge from medical reports. These reports are often maintained as unstructured free-text and constitute a large volume of data. Knowledge and information extraction using Natural Language Processing (NLP) techniques could therefore help improve health care and medical decision-making. This stage is often preceded by a text segmentation phase to identify the part of text that are of great interest. In this article, we present our automatic segmentation system of medical reports into predefined categories and which consists of two complementary steps. First, an algotithm based on titles detection for the identification of certain sections. The second part is a supervised sentence classification task based on deep learning algorithms. This system was evaluated on 500 reports and achieved more than 96% classification accuracy.