Participation de l'équipe LAI à DEFT 2019

Jacques Hilbey[1] Louise Deléger[2] Xavier Tannier[3]
(1) Inserm, LIMICS
(2) MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France
(3) Sorbonne Université, Inserm, LIMICS, Paris, France
jacques.hilbey@inserm.fr, louise.deleger@inra.fr,
xavier.tannier@sorbonne-universite.fr

RÉSUMÉ

Nous présentons dans cet article les méthodes conçues et les résultats obtenus lors de notre participation à la tâche 3 de la campagne d'évaluation DEFT 2019. Nous avons utilisé des approches simples à base de règles ou d'apprentissage automatique, et si nos résultats sont très bons sur les informations simples à extraire comme l'âge et le sexe du patient, ils restent mitigés sur les tâches plus difficiles.

ABSTRACT

Participation of team LAI in the DEFT 2019 challenge

We present in this article the methods developed and the results obtained during our participation in task 3 of the DEFT 2019 evaluation campaign. We used simple rule-based or machine-learning approaches ; our results are very good on the information that is simple to extract (age, gender), they remain mixed on the more difficult tasks.

1 Introduction

Nous présentons dans cet article les méthodes que nous avons conçues pour répondre à la tâche 3 de la campagne d'évaluation DEFT 2019. Cette tâche porte sur l'extraction d'informations dans un corpus de cas cliniques (Grabar *et al.*, 2018). Quatre types d'information doivent être identifiés :

- l'âge du patient ;
- le genre du patient ;
- l'origine : le motif de la consultation ou de l'hospitalisation ;
- l'issue, à déterminer parmi cinq valeurs possibles : guérison, amélioration, stable, détérioration ou décès.

Le jeu d'entraînement fourni par les organisateurs est composé de 290 documents, tandis que le jeu de test est composé de 427 cas. Plus de détails sur le défi sont présents dans Grabar *et al.* (2019).

2 Méthode

Nous avons mis en place des méthodes à base de règles et/ou d'apprentissage automatique, selon le type d'information à extraire.

2.1 Âge et genre

En étudiant le corpus d'entraînement, nous avons constaté que les informations d'âge et de sexe étaient souvent exprimées sous des formes similaires avec relativement peu de variations dans les documents. Ceci a orienté notre choix de méthode vers une approche à base de règles. Il nous a en effet semblé que la majorité de ces informations pourrait être capturée à l'aide d'un ensemble relativement restreint de règles.

Nous avons implémenté nos règles à l'aide de l'outil PyRATA (Hernandez & Hazem, 2018). PyRATA est une librairie python qui permet d'écrire des règles de type expressions régulières s'appliquant à des structures de données plus complexes qu'une chaîne de caractères, en particulier sur des suites de tokens possédant différents attributs (lemmes, étiquettes morpho-syntaxiques, etc.).

Pour l'âge, nous avons conçu un ensemble de motifs composés soit de déclencheurs (*âgé de, patient de, etc.*) suivi d'un nombre (en chiffre ou en lettre) et d'une unité de temps, soit de noms dénotant des classes d'âge (*quadragénaire, quinquagénaire, etc.*). Les âges exprimés en mois ont été convertis en années (en pratique, 0 ou 1).

Pour le genre, nous avons établi des listes d'expressions évocatrices du genre masculin (*homme, patient, Monsieur, testicule, un enfant, prénom masculin, etc.*) et du genre féminin (*femme, patiente, Madame, utérus, une enfant, prénom féminin, etc.*).

Dans les rares cas où plusieurs personnes sont concernées par le cas, nous nous sommes assurés de la cohérence entre le nombre d'âges et de genres retournés, en considérant le plus petit nombre.

Avant d'appliquer nos règles, nous pré-traitons le corpus (segmentation en mots, analyse morphosyntaxique) à l'aide de la librairie python spaCy [1].

2.2 Origine

Comme pour l'âge et le sexe, nous avons adopté une approche à base de règles, implémentées à l'aide de l'outil pyRATA, pour capturer l'information relative à l'origine de l'admission.

Celle-ci se trouve le plus souvent, sans surprise, dans la première ou la deuxième phrase du cas. Un premier motif est apparu : une portion de phrase ayant pour introducteur la préposition *pour* et pour terminateur le point final de la phrase. Nous avons procédé ensuite par ajout d'autres introducteurs possibles liés soit au patient (*présenter, souffrir, subir, se plaindre, etc.*) soit au processus médical (*prise en charge, diagnostic, tableau, bilan, etc.*).

Dans un deuxième temps, nous avons étendu les terminateurs possibles à d'autres ponctuations et préféré éventuellement, au simple *pour*, des expressions plus spécifiques (*hospitaliser / admettre / consulter pour*).

2.3 Issue

L'issue caractérise l'évolution de l'état clinique du patient entre son admission et la fin de la consultation ou de l'hospitalisation. Nous avons suivi deux voies différentes pour capter cette évolution.

1. `https://spacy.io/`

Une première méthode (*run 1*) a consisté à utiliser des tests du χ^2 afin de déterminer quels N-grammes (avec N de 1 à 3), dans les cas cliniques présentant une même issue, lui étaient le plus fortement associés, puis à sélectionner manuellement pour chaque issue une dizaine de ces suites de mots en évitant autant que possible ceux qui paraissaient trop caractéristiques du jeu d'entraînement. En recherchant dans chaque cas clinique ces N-grammes, il a été possible d'établir pour chacun un score par issue et d'associer au cas l'issue ayant le score maximal. En cas de score nul, l'issue attribuée était 'NUL' et en cas d'égalité entre plusieurs issues, la plus fréquente de celles-ci dans le jeu d'entraînement.

La deuxième méthode (*run 2*) a consisté à décrire la deuxième moitié de chaque cas clinique (où les informations relatives à l'issue étaient le plus susceptibles de se trouver) dans un espace vectoriel de type *one-hot vectors* (sac de mots), avec une pondération TF-IDF, puis à appliquer aux vecteurs ainsi constitués des classifieurs multi-classe courants (régression logistique, bayésien naïf multinomial, séparateur à vaste marge linéaire, forêt aléatoire, arbre de décision, k plus proches voisins) dont nous avons fait varier les paramètres pour choisir finalement la meilleure configuration sur le jeu d'entraînement, par validation croisée.

Le cas de plusieurs issues, rarissime dans le corpus d'entraînement, a été ignoré.

	Précision	Rappel
Âge	0.986	0.953
Genre	0.993	0.983
Issue (*run 1*)	0.693	0.619
Issue (*run 2*)	Exactitude (*Accuracy*)	
	0.524	

TABLE 1 – Performances pour l'âge, le genre et l'issue sur les données d'entraînement (pour le *run 2*, performance en validation croisée utilisée pour le choix du meilleur classifieur).

	Précision	Rappel	F-Mesure	Meilleure F-mesure
Âge	0.980	0.919	0.948	0.948
Genre	0.981	0.974	0.978	0.978
Issue (*run 1*)	0.486	0.405	0.442	0.505
Issue (*run 2*)	0.498	0.492	0.495	

TABLE 2 – Performances pour l'âge, le genre et l'issue sur les données de test (résultats officiels fournis par les organisateurs du challenge), comparées aux meilleurs résultats parmi les participants.

3 Résultats et discussion

Les règles de reconnaissance de l'âge et du genre du patient donnent de très bons résultats sur le corpus d'entraînement (tableau 1). Les performances sur le corpus de test sont également bonnes mais légèrement en baisse (tableau 2), en particulier pour l'âge qui perd environ 3 points en rappel. Nos règles semblent donc dans l'ensemble assez robustes, mais manquent certains cas.

Dans les mêmes tableaux, il est intéressant mais pas surprenant de constater que l'approche à base de règles pour l'issue connaît une forte baisse de performance entre le jeu d'entraînement et le jeu de

Origine	Macro-	Précision	0.582
		Rappel	0.722
		F-mesure	0.645
		Meilleure F-mesure	0.666
	Micro-	Précision	0.628
		Rappel	0.735
		F-mesure	0.677
		Meilleure F-mesure	0.677
		overlap-accuracy	0.600

TABLE 3 – Performances pour l'origine (admission) sur les données de test (résultats officiels fournis par les organisateurs du challenge), comparées aux meilleurs résultats parmi les participants.

test. L'approche par apprentissage est en revanche plus robuste (le classifieur finalement retenu a été un séparateur à vaste marge (SVM) linéaire avec les paramètres $C = 10^{-6}$ et $tol = 10^{-6}$) et conduit à notre meilleur résultat sur le test. Dans les deux cas, la difficulté de la tâche cumulée au faible nombre de documents d'entraînement ne permet pas d'obtenir des résultats satisfaisants avec ces approches simples.

Les matrices de confusion pour l'entraînement (Figure 1) et pour le test (Figure 2) permettent de mettre en évidence la difficulté pour nos deux approches de distinguer efficacement les classes *amélioration* et *guérison*, ce qui représente la plus grande partie des erreurs des systèmes. Il serait intéressant de savoir si les chiffres d'accords inter-annotateurs du corpus illustrent les mêmes difficultés.

Enfin, la table 3 présente les résultats obtenus pour l'extraction du motif de l'admission (origine) avec les métriques officielles du défi.

4 Conclusion

Si nos résultats sont, selon les cas, les meilleurs ou tout proches des meilleurs du challenge, nous ne sommes malheureusement pas parvenus à produire des approches innovantes et performantes sur cette tâche du défi DEFT 2019. Nos résultats proviennent d'approches classiques et conduisent à des résultats sans surprise : très bons sur les informations simples à extraire comme l'âge et le sexe, plus mitigés sur les informations pouvant s'exprimer de façon très variable dans les textes.

Nous avons également réalisé des expérimentations sur la tâche 1 (extraction de mots-clés), avec une extraction de termes puis une adaptation de la Word Mover Distance (Kusner *et al.*, 2015) pondérée par les tf.idf des termes, pour estimer les termes les plus représentatifs de chaque paire cas/discussion, mais les résultats se sont avérés peu convaincants.

Remerciements

Nous remercions les organisateurs pour la création du corpus ainsi que pour l'organisation du défi, ainsi qu'Ivan Lerner pour ses expériences préliminaires.

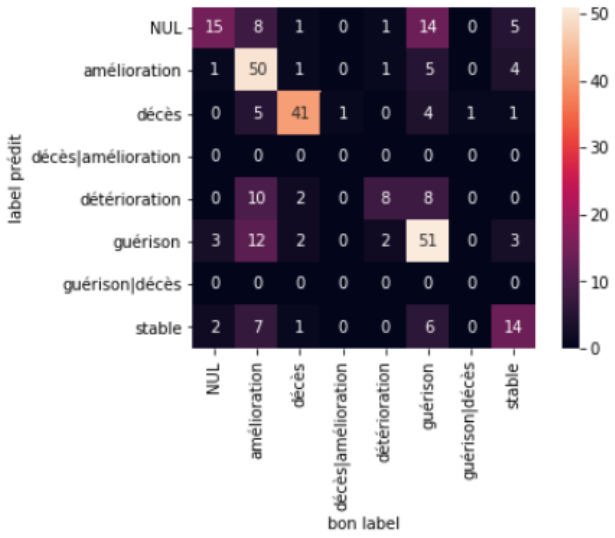
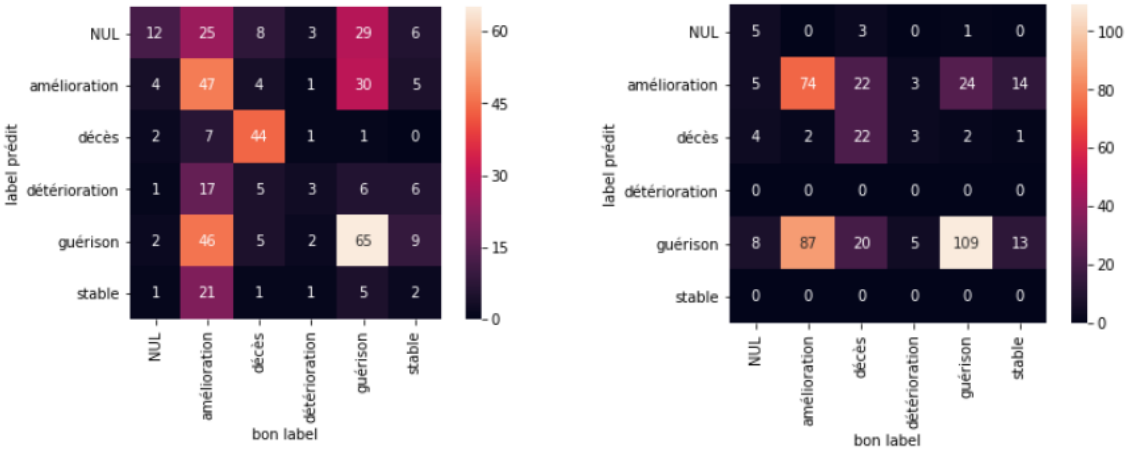


FIGURE 1 – Matrice de confusion pour l'*issue (run 1)* sur le jeu d'entraînement.



run 1 *run 2*

FIGURE 2 – Matrice de confusion pour l'*issue* sur le jeu de test (run 1 à base de règles et run 2 par apprentissage).

Références

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). CAS : French corpus with clinical cases. In *LOUHI 2018 : The Ninth International Workshop on Health Text Mining and Information Analysis*.

GRABAR N., GROUIN C., HAMON T. & CLAVEAU V. (2019). Recherche et extraction d'information dans des cas cliniques. présentation de la campagne d'évaluation deft 2019. In *Actes de DEFT*.

HERNANDEZ N. & HAZEM A. (2018). PyRATA, Python Rule-based feAture sTructure Analysis. In N. C. C. CHAIR), K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, K. HASIDA, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS & T. TOKUNAGA, Eds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France : European Language Resources Association (ELRA).

KUSNER M. J., SUN Y., KOLKIN N. I. & WEINBERGER K. Q. (2015). From Word Embeddings To Document Distances. In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, Lille, France.