HAL
open science

Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques

Rémi Cardon, Natalia Grabar, Cyril Grouin, Thierry Hamon

To cite this version:

Rémi Cardon, Natalia Grabar, Cyril Grouin, Thierry Hamon. Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques. 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition), 2020, Nancy, France. pp.1-13. hal-02784737v2

HAL Id: hal-02784737

https://hal.science/hal-02784737v2

Submitted on 17 Jun 2020 (v2), last revised 23 Jun 2020 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques

Rémi Cardon[1, 2] Natalia Grabar[1, 2] Cyril Grouin[3] Thierry Hamon[3, 4]
(1) CNRS, UMR 8163, F-59000 Lille, France
(2) Univ. Lille, UMR 8163 – STL – Savoirs Textes Langage, F-59000 Lille, France
(3) Université Paris-Saclay, CNRS, LIMSI, F-91400 Orsay, France
(4) Université Paris 13, Sorbonne Paris Cité, F-93430, Villetaneuse, France
prenom.nom@univ-lille.fr, prenom.nom@limsi.fr

Résumé

L'édition 2020 du défi fouille de texte (DEFT) a proposé deux tâches autour de la similarité textuelle et une tâche d'extraction d'information. La première tâche vise à identifier le degré de similarité entre paires de phrases sur une échelle de 0 (le moins similaire) à 5 (le plus similaire). Les résultats varient de 0,65 à 0,82 d'EDRM. La deuxième tâche consiste à déterminer la phrase la plus proche d'une phrase source parmi trois phrases cibles fournies, avec des résultats très élevés, variant de 0,94 à 0,99 de précision. Ces deux tâches reposent sur un corpus du domaine général et de santé. La troisième tâche propose d'extraire dix catégories d'informations du domaine médical depuis le corpus de cas cliniques de DEFT 2019. Les résultats varient de 0,07 à 0,66 de F-mesure globale pour la sous-tâche des pathologies et signes ou symptômes, et de 0,14 à 0,76 pour la sous-tâche sur huit catégories médicales. Les méthodes utilisées reposent sur des CRF et des réseaux de neurones.

Abstract

Presentation of the DEFT 2020 Challenge : open domain textual similarity and precise information extraction from clinical cases

The 2020 edition of the French Text Mining Challenge proposed two tasks about textual similarity and one information extraction task. The first task aims at identifying the degree of similarity between pairs of sentences, from 0 (the less similar) to 5 (the most similar). The results range from 0.65 to 0.82 (EDRM). The second task consists in identifying the closest sentence from a source sentence among three given sentences. The results are very high, ranging from 0.94 to 0.99 (precision). Both tasks rely on a corpus from the general and health domains. The third task proposes to extract ten categories of information from the medical domain, from a corpus of clinical cases used during the last competition. The results range from 0.07 to 0.66 (F-measure) on the sub-task identifying pathologies and signs or symptoms, and from 0.14 to 0.76 for the sub-task concerning eight medical categories. Methods used rely on CRF and neural networks.

Mots-clés : Cas cliniques, extraction d'information, similarité textuelle.

Keywords: Clinical Cases ; Information Extraction ; Textual Similarity.

Atelier DÉfi Fouille de Textes, pages 1–13.
HAL : hal-02784737.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

1 Introduction

Cet article présente la campagne d'évaluation 2020 du défi fouille de texte (DEFT). Cette nouvelle édition se compose de trois tâches, dont deux nouvelles portant sur la similarité entre phrases et l'identification de phrases parallèles en domaine général et domaine de spécialité, tandis que la troisième tâche reprend le corpus de cas cliniques utilisé pour DEFT 2019 (Grabar *et al.*, 2019) et propose de travailler sur des catégories d'informations plus fines.

La similarité textuelle offre plusieurs possibilités d'utilisation, de la détection du plagiat jusqu'à la réécriture dans un but de simplification. Dans cette campagne, nous avons souhaité étudier les niveaux de similarité qu'il est possible d'inférer automatiquement, entre phrases portant soit sur le même sujet, soit des sujets fondamentalement différents.

L'extraction d'informations précises contenues dans des documents de spécialité, tels que les cas cliniques, constitue une première étape pour l'accès à l'information, la recherche de cas similaires, ou encore le peuplement de bases de données. A l'image de la campagne d'évaluation américaine i2b2/VA 2010, nous avons proposé pour le français une tâche d'extraction d'informations pour le domaine médical.

Nous avons lancé le défi le 27 janvier. Treize équipes se sont inscrites. Dix équipes ont participé à la phase de test, qui s'est déroulée du 28 au 30 avril. Parmi les équipes étant allé au terme de la campagne, nous retenons la présence d'une équipe académique africaine (Université Assane Seck Ziguinchor, Sénégal) et de cinq équipes industrielles (EDF R&D, Palaiseau ; Reezocar, Boulogne-Billancourt ; Synapse, Toulouse) ou mixtes industrielles et académiques (IBM France, Bois-Colombes et Université Paris 7, Paris ; LIRMM et ONAOS, Montpellier). Quatre équipes académiques complètent la liste des participants (BiTeM, Haute Ecole Spécialisée, Genève ; DOING, Université d'Orléans ; LIMICS, Sorbonne Université, Paris ; LIPN, Sorbonne Paris-Nord, Villetaneuse). La diffusion des corpus d'entraînement s'est faite à partir du 3 février pour la première tâche, à partir du 13 février pour la troisième tâche et du 17 février pour la deuxième tâche, pour des raisons de finalisation des corpus. Enfin, une version améliorée des annotations de la troisième tâche a été distribuée le 20 mars.

2 Présentation

Le corpus utilisé pour les deux premières tâches se compose de pages Wikipedia et Vikidia [1] relatives à différents sujets (par exemple, les pages Almaty, Apiculture, Biberon, Boris Godounov, etc.) ainsi que du contenu en santé tel que des notices de médicaments (Bromazepam, Buprénorphine, etc.) et des résumés Cochrane (Grabar & Cardon, 2018).

2.1 Tâche 1 – Degré de similarité entre paires de phrases

La première tâche consiste à identifier le degré de similarité entre deux phrases, sur une échelle de valeurs comprises entre 0 (le moins similaire) et 5 (le plus similaire), sans que la sémantique associée à chaque valeur de cette échelle n'ait été définie. Cinq personnes ont annoté les paires de phrases du corpus, chacune ayant son interprétation personnelle du type de contenu associé à chaque degré dans

1. Vikidia est la Wikipedia destinée aux 8–13 ans, https://fr.vikidia.org/wiki/Vikidia:Accueil

la mesure où nous n'avons pas souhaité fournir de définition des degrés de similarité. Pour constituer les valeurs de référence, nous avons retenu la valeur issue du vote majoritaire (Cardon & Grabar, 2020). Le corpus d'entraînement intègre 600 paires de phrases tandis que le corpus d'évaluation contient 410 paires. Le tableau 1 fournit quelques exemples de paires de phrases source et cible avec le degré de similarité issu de l'annotation humaine.

Phrases source et cible	Degré
Il commence par s'intéresser à la résistance à la faim, la soif et à la fatigue en 1951.	0
Pour prouver qu'on pouvait vivre sans eau ni nourriture, il traversa en solitaire l'Atlantique sans autres ressources que les poissons, le plancton, l'eau de pluie et de petites quantités d'eau de mer durant 65 jours.	
En cas de survenue d'une hypotension importante, le patient doit être mis en décubitus dorsal, et recevoir, si nécessaire, une perfusion iv de chlorure de sodium.	1
Si une hypotension importante se produit, elle peut être combattue en allongeant le patient jambes relevées.	
Deux essais (106 participants) comparaient l'héparine de bas poids moléculaire à un placebo ou à l'absence de traitement.	2
Deux essais (259 participants) comparaient l'héparine à l'absence de traitement.	
La vieille ville est entourée de remparts, érigés au XIIIe siècle, très appréciés par les promeneurs.	3
La ville haute, ceinte de remparts, est très pittoresque.	
Le biberon, (du latin bibere, « boire ») ou bouteille en Suisse, est un ustensile utilisé pour l'allaitement artificiel.	4
Un biberon (une bouteille, en Suisse), est un outil permettant d'allaiter un bébé artificiellement, ou naturellement, si la mère a tiré son lait.	
Le médecin spécialisé pratiquant la neurologie s'appelle le neurologue.	5
Le médecin qui pratique la neurologie est le neurologue.	

TABLE 1 – Degré de similarité pour quelques paires de phrases source et cible de la tâche 1

Le tableau 2 présente le nombre et le pourcentage d'annotations pour chaque degré de similarité dans les corpus d'entraînement et d'évaluation. On observe que le degré de similarité le plus faible (0) est celui qui contient le plus d'annotations dans les deux corpus (plus du tiers du nombre total d'annotations), suivi de l'avant-dernier degré (4), couvrant déjà plus de la moitié des paires du corpus.

Corpus	Degrés de similarité											
	0		1		2		3		4		5	
Entraînement (600 paires)	216	36,0%	56	9,3%	29	4,8%	66	11,0%	136	22,7%	97	16,2%
Evaluation (410 paires)	147	35,9%	37	9,0%	28	6,8%	44	10,7%	90	22,0%	64	15,6%

TABLE 2 – Nombre et pourcentage d'annotations par degré de similarité dans les corpus de la tâche 1

Bien qu'aucune définition n'ait été fournie, les annotateurs humains ont sensiblement convergé vers

les observations suivantes : degré 5 pour une similarité quasi parfaite, degré 4 si l'une des deux phrases apporte une information de plus, degré 3 si une information importante est manquante, degrés 2 ou 1 en fonction du niveau de reformulation, et degré 0 pour une absence de similarité ou trop complexe. Nous observons que les participants du défi ont eu une interprétation similaire de ces degrés de similarité, comme celle indiquée par Cao *et al.* (2020) pour l'équipe EDF R&D.

2.2 Tâche 2 – Identification des phrases parallèles

La deuxième tâche vise à identifier, parmi trois phrases cibles, celle qui correspond le mieux à la phrase source en terme de phrase parallèle. Une réponse parmi les trois phrases cibles fournies est toujours attendue, les participants ont donc l'obligation d'identifier une phrase parallèle pour chaque ensemble de phrases source et cibles. Le parallélisme des phrases est lié à la relation simple-compliqué : la phrase source correspond au contenu compliqué alors que les phrases simples contiennent le contenu simple ou simplifié. L'une des phrases simples est donc dérivée de la phrase compliquée. La tableau 3 fournit des exemples de phrases source et cibles sur différents sujets. On observe ainsi que, dans certains cas, la même phrase a été utilisée comme phrase source et comme l'une des phrases cibles (deuxième exemple), et que dans d'autres cas, des indices numériques tels que les dates ne correspondent pas forcément (troisième exemple). Le corpus d'entraînement comprend 572 ensembles de phrases source et cibles tandis que le corpus d'évaluation en compte 530.

Type	Phrases proposées
Source	Arrivé en France en 1972, ce chat reste méconnu en dehors de son pays d'origine.
Cibles	Les principaux matériaux sont le grès et la latérite.
	Ce chat est apparu dans une portée d'American Shorthairs, en 1966, dans l'État de New-York.
	Bien qu'il soit apparu en France dès 1972, ce chat reste méconnu hors de son État d'origine.
Source	en suède, le taux légal est de 0,2 g par litre de sang
Cibles	en suisse, le taux légal est de 0,5 g/l de sang ou 0,22 mg d'alcool par litre d'air expiré, depuis 2005
	en suède, le taux légal est de 0,2 g par litre de sang
	en belgique, le taux légal est de 0,5 g/l de sang ou 0,22 mg d'alcool par litre d'air expiré
Source	En 1534, il est appelé comme maître d'œuvre par le comte Giangiorgio Trissino pour diriger le chantier de la villa Cricoli.
Cibles	*En 1537, il est appelé par le comte Giangiorgio Trissino pour diriger le chantier de la villa Cricoli.*
	Trissino est un humaniste, poète, philosophe et diplomate au service de la curie romaine (le gouvernement pontifical) ; c'est aussi un passionné d'architecture .
	Le théâtre Olympique, achevé après 1580, est l'œuvre ultime de Palladio, terminée après sa mort par son fils Silla et son disciple Scamozzi.

TABLE 3 – Exemples de phrases sources et cibles pour la tâche 2. La phrase cible la plus parallèle de la phrase source apparaît en italiques

2.3 Tâche 3 – Extraction d'information fine

Présentation Dans la continuité de DEFT 2019, qui portait sur l'analyse de cas cliniques, une sous-partie du corpus utilisé [2] (Grabar *et al.*, 2018) a été annotée avec des catégories d'informations médicales fines autour de quatre domaines [3] : autour des patients *(anatomie)*, de la pratique clinique *(examen, pathologie, signe ou symptôme)*, des traitements médicamenteux et chirurgicaux *(dose, durée, fréquence, mode d'administration, substance, traitement, valeur)*, et autour du temps *(date, moment)* (Grouin *et al.*, 2019). Les annotations de pathologies et de signes ou symptômes couvrent aussi bien des mots isolés que de longues portions textuelles (jusqu'à 33 mots dans une seule portion [4]), alors que les annotations des autres catégories sont généralement plus courtes (entre un et quatre mots). De plus, les annotations d'examens, de pathologies et de signes ou symptômes peuvent englober des annotations d'autres catégories (telles que des parties anatomiques ou des valeurs numériques). Nous avons complété ces annotations avec l'information sur les assertions des pathologies et signes ou symptômes *(présent, absent, possible, hypothétique, non-associé)*, de la norme des valeurs d'examens biologiques et physiques *(normal, haut, bas)* et de la prise des traitements et substances *(arrêt, reprise)*. Ces catégories d'annotation s'inspirent de celles utilisées en 2010 dans la campagne d'évaluation internationale i2b2/VA (Uzuner *et al.*, 2011). Seules dix catégories [5] d'annotations ont été utilisées dans cette campagne. Nous présentons un extrait de corpus annoté sur la figure 1.

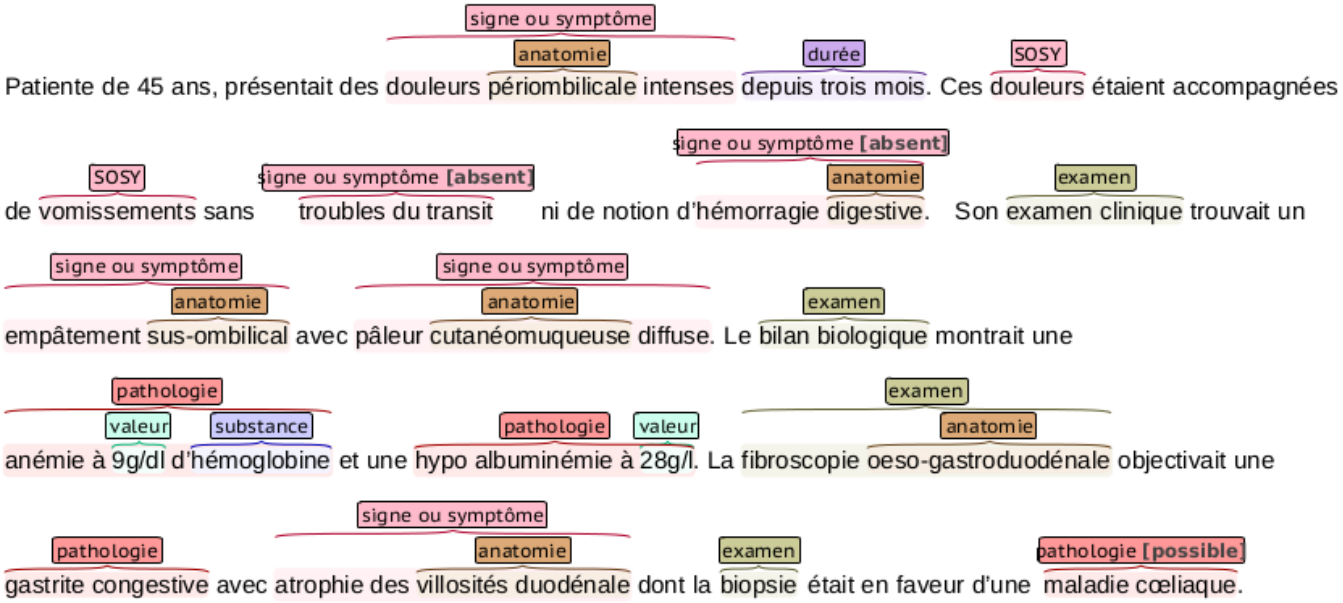

FIGURE 1 – Extrait de corpus annoté. Les boîtes renvoient aux catégories, les éléments entre crochets aux informations complémentaires

Le corpus d'entraînement comprend 100 fichiers (soit 7844 annotations sur les dix catégories) tandis que le corpus d'évaluation compte 67 fichiers (4738 annotations). Le tableau 4 renseigne du nombre et du pourcentage d'annotations par catégorie dans les corpus d'entraînement et d'évaluation, pour

2. Nous avons aléatoirement extrait 167 fichiers parmi les 717 fichiers du corpus utilisé l'année dernière.
3. Le guide d'annotation est accessible à l'adresse : `https://deft.limsi.fr/2020/guide-deft.html`
4. La portion suivante a intégralement été annotée comme un signe ou symptôme : « *volumineuse masse pelvienne de 170 sur 150 x 120 mm, de contours polylobés, de densité spontanément hétérogène, se rehaussant de façon intense et hétérogène après injection, avec des zones d'hypodensité, quasi liquidiennes* » car renvoyant à une description complète d'une observation.
5. Anatomie, dose, examen, mode, moment, pathologie, signe ou symptôme, substance, traitement, valeur.

les catégories utilisées dans cette édition de la campagne DEFT.

Catégories utilisées	Corpus			
	Entraînement (100 fichiers)		Evaluation (67 fichiers)	
Anatomie	1454	18,5%	1121	23,7%
Dose	380	4,9%	52	1,1%
Examen	1219	15,5%	817	17,2%
Mode	225	2,9%	89	1,9%
Moment	453	5,8%	165	3,5%
Pathologie	368	4,7%	166	3,5%
Signe ou symptôme	1799	22,9%	1279	27,0%
Substance	1016	13,0%	313	6,6%
Traitement	383	4,9%	304	6,4%
Valeur	547	7,0%	432	9,1%

TABLE 4 – Nombre et pourcentages d'annotations par catégorie dans les corpus de la tâche 3

Annotation humaine Deux annotateurs ont réalisé ce travail en utilisant l'interface d'annotation BRAT (Stenetorp *et al.*, 2012). Compte tenu de l'ampleur du travail d'annotation à réaliser (au total, 8615 entités seront annotées dans le corpus d'entraînement et 4933 dans le corpus d'évaluation), nous avons opté pour une annotation séquentielle du corpus d'entraînement. D'abord composé de 70 fichiers, ce corpus a été annoté par un premier annotateur puis corrigé par le deuxième annotateur. Trente fichiers supplémentaires ont été annotés par le deuxième annotateur, puis corrigés par le premier annotateur. En conséquence, il n'est pas possible de calculer d'accord inter-annotateur sur les fichiers du corpus d'entraînement. Néanmoins, nous avons mesuré l'écart avant et après correction, et les résultats obtenus [6] nous ont confortés dans la possibilité de poursuivre avec les choix effectués, en ajustant les règles du guide d'annotation au besoin.

L'annotation des 67 fichiers du corpus d'évaluation s'est faite sur la base d'une pré-annotation automatique, réalisée grâce à un modèle CRF entraîné sur les cent fichiers du corpus d'entraînement. Les annotateurs humains ont corrigé cette pré-annotation de manière indépendante, puis réalisé une phase d'adjudication. Malgré la pré-annotation automatique, les accords inter-annotateur calculés sur ces fichiers se montent à 0,67 de F-mesure en évaluation stricte et 0,80 en évaluation souple [7] sur les treize catégories annotées par les humains. Les catégories pour lesquelles les humains ont obtenu de moins bonnes performances sont les pathologies, les signes ou symptômes, ainsi que les fréquences et modes d'administration. Les différences observées concernent des choix différents de catégories ainsi que des oublis. Nous observons également que les deux annotateurs ont pu choisir des mots différents pour annoter le même concept [8]. Enfin, l'absence de formation médicale des annotateurs peut également présenter un obstacle dans la qualité du travail d'annotation. Par exemple, la différence entre pathologie et signe ou symptôme reste complexe. Les annotateurs ont notamment

6. Nous calculons une F-mesure stricte globale de 0,8183 et une F-mesure souple de 0,8946 sur les 70 premiers fichiers, et une F-mesure stricte de 0,9565 et une F-mesure souple de 0,9781 sur les trente derniers fichiers.

7. L'évaluation stricte repose sur un appariement exact entre frontières et étiquette. Une évaluation souple accepte quelques caractères d'écart au niveau des frontières.

8. Pour le mode d'administration, un annotateur aura annoté le mot *injecte* alors que l'autre aura annoté le mot *seringue* de la même phrase : « L'infirmière anesthésiste prépare la seringue de morphine et la remet à l'anesthésiste qui l'injecte. »

considéré que les tumeurs bénignes sont des signes ou symptômes, alors que les tumeurs malignes sont des pathologies. Certains mots ou suffixes sont également des indices importants pour déterminer la catégorie [9]. Concernant la taille des portions, les annotateurs ont établi qu'une portion doit contenir le maximum d'informations se rapportant à la même idée. Ce choix explique la taille de certaines portions annotées en signes ou symptômes.

Une fois ce travail d'annotation terminé, nous sommes revenus sur les annotations du corpus d'entraînement pour le rendre plus homogène [10] avec le corpus d'évaluation. Cette deuxième version du corpus d'entraînement a été distribuée aux participants le 20 mars.

3 Evaluation

Tâche 1 Les degrés de similarité de la tâche 1 renvoient à une graduation sur une échelle de six valeurs. Nous avons donc retenu comme mesure principale la distance relative moyenne à la solution, calculée en micro-moyenne. A chacune des six valeurs possibles pour la donnée de référence r_i, correspond une valeur de distance maximale possible entre la réponse du système et cette donnée $dmax(h_i, r_i)$. Ainsi, si le degré de similarité attendu est 5, la distance maximale possible est alors maximale ($5 - 0 = 5$), mais si le degré attendu est 2, la distance maximale possible sera de 3 ($5 - 2 = 3$ alors que $2 - 0 = 3$). L'exactitude en distance relative à la solution moyenne (EDRM) se calcule en micro-moyenne comme indiqué dans l'équation 1.

$$\text{EDRM} = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{d(h_i, r_i)}{dmax(h_i, r_i)}\right) \tag{1}$$

Tâche 2 Sur cette tâche, il est possible d'évaluer, soit la meilleure réponse ramenée par le système (la phrase cible la plus similaire à la phrase source), soit le classement ou l'ordre des trois phrases cibles proposées, qui auront été ramenées de la plus similaire à la moins similaire. Pour évaluer un classement, la moyenne des précisions non interpolées $P(I_i^j)$ calculées à chaque position, dans la liste des hypothèses, d'une des n_i réponses correctes I_i^j pour la phrase source S_i, est alors une mesure pertinente. Sur l'ensemble des phrases source et cible, nous utilisons alors la moyenne de cette précision moyenne (MAP, voir formule 2). Dans le cas où un système ne classerait pas toutes les phrases cibles fournies, et notamment la phrase cible attendue, sa précision moyenne est alors nulle, comme si elle avait été classée à l'infini par le système.

$$\text{MAP} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i}\sum_{j=1}^{n_i}P(I_i^j) \tag{2}$$

Si une seule phrase cible est renvoyée pour chaque phrase source, une précision classique est alors employée.

9. Le suffixe *-mégalie* qui renvoie à un gonflement *(hépatomégalie, splénomégalie)* permet d'annoter un signe ou symptôme, tandis que les *(adéno)carcinomes* et *adénopathies* seront annotés en pathologie.

10. Cette nouvelle version s'accompagne d'un nombre plus élevé d'annotations dans les catégories *anatomie* (de 1454 à 1608), *signe ou symptôme* (de 1799 à 1831) et *valeurs* (de 547 à 588) pour les plus fortes hausses. Dans le même temps, certaines annotations de la catégorie *traitement* ont été supprimées (passant de 383 à 374 annotations).

Tâche 3 La dernière tâche est une tâche de repérage d'entités nommées. Les mesures habituelles de précision, rappel et F-mesure, calculées sur chaque catégorie et au niveau global, constituent le mode d'évaluation le plus pertinent et le mieux compris du point de vue des valeurs calculées.

4 Résultats

4.1 Tâche 1 – Degré de similarité entre paires de phrases

Le tableau 5 présente les résultats en EDRM (mesure officielle) ainsi que la corrélation de Spearman et la p-value sur les prédictions de la première tâche. Sur l'ensemble des soumissions, la moyenne est de 0,7617 et la médiane se situe à 0,7947.

Soumission	EDRM	Corrélation de Spearman	p-value
EDF R&D, 1	0,8198	0,7305	1,3963e-69
EDF R&D, 2	0,8018	0,7105	2,9216e-64
EDF R&D, 3	0,8069	0,7164	9,2243e-66
Reezocar, 1	0,7919	0,7060	4,1583e-63
Reezocar, 2	0,8105	0,7352	6,6003e-71
Reezocar, 3	0,8022	0,7080	1,2883e-63
Sorbonne, 1	0,7092	0,7485	8,4089e-75
Sorbonne, 2	0,6734	0,7321	5,2500e-70
Sorbonne, 3	0,8147	0,7479	1,2845e-74
Synapse, 1	0,6533	0,7499	3,1295e-75
Synapse, 2	0,6663	0,7421	7,0960e-73
Synapse, 3	0,6838	0,7679	6,1899e-81
UASZ, 1	0,7947	0,7528	4,3371e-76
UASZ, 2	**0,8217**	0,7691	2,3769e-81
UASZ, 3	0,7755	0,7769	5,5766e-84

TABLE 5 – Evaluation des prédictions en EDRM. Le meilleur résultat est en gras

Nous avons utilisé le test de Student pour évaluer la significativité statistique des résultats des participants à la tâche 1 (voir figure 2). De manière globale, nous observons que les résultats des équipes EDF R&D, Sorbonne et REEZOCAR n'ont pas de différence significative entre eux, tandis que ceux de l'équipe Synapse diffèrent statistiquement de ceux de tous les autres participants. Aussi, les résultats de l'équie UASZ sont significativement différents de la deuxième soumission de l'équipe Sorbonne et des soumissions de l'équipe REEZOCAR. Enfin, la troisième soumission de l'équipe UASZ est éloignée de l'ensemble des soumissions. De plus, nous pouvons remarquer plusieurs spécificités des soumissions. La meilleure soumission (UASZ_2) n'est pas significativement différente des résultats des équipes EDF R&D, Sorbonne (à l'exception de la deuxième soumission), et de la troisième soumission de l'équipe REEZOCAR. De même, il n'y a pas de différences significatives entre les soumissions d'une même équipe, à l'exception de celles de Synapse. Ceci peut s'expliquer par le fait que les soumissions sont toutes basées sur la même méthode et ne diffèrent qu'à travers des paramétrages particuliers. En revanche, les variations significatives dans les soumissions (1 et 3 vs. 2) de Synapse peuvent certainement s'expliquer par l'utilisation des modèles BERT ou MUSE.

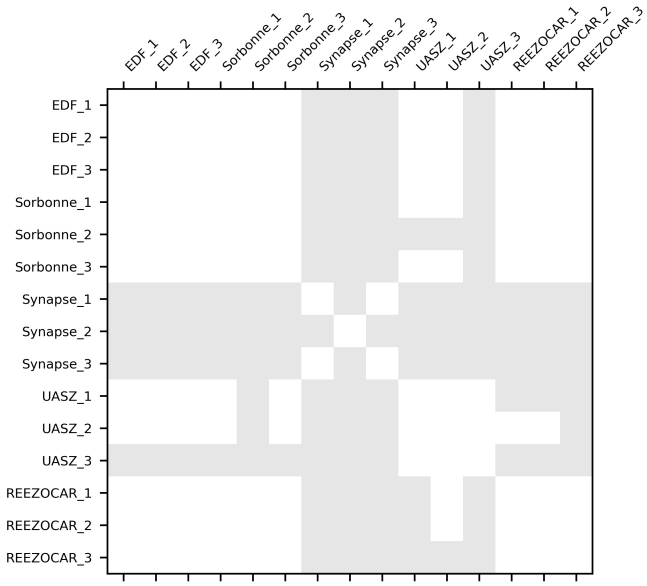

FIGURE 2 – Significativité statistique des soumissions de la tâche 1. Les zones grisées représentent une différence significative entre deux soumissions.

Méthodes Sur cette tâche, les méthodes utilisées par les participants sont variées. Plusieurs reposent sur des calculs de distance entre phrases (distances euclidiennes, Jaccard, Manhattan), avec des représentations de type TF×IDF, notamment par l'équipe Sorbonne (Buscaldi *et al.*, 2020) qui a mesuré l'intérêt et l'absence d'intérêt de cette représentation, dont ces distances peuvent ensuite servir de traits pour des algorithmes d'apprentissage (régression logistique, forêt d'arbres, graphes sémantiques, etc.), approche qui a été suivie par les équipes EDF R&D (Cao *et al.*, 2020), Reezocar (Tapi Nzali, 2020) et UASZ (Drame *et al.*, 2020). Cette dernière équipe obtient les meilleures performances de la tâche avec sa deuxième soumission fondée sur un perceptron multi-couche. D'autres participants ont utilisés des modèles de plongements lexicaux multilingues dérivés de BERT, en particulier Sentence M-BERT et MUSE par Synapse (Belkacem *et al.*, 2020).

4.2 Tâche 2 – Identification des phrases parallèles

Le tableau 6 présente les résultats sur la deuxième tâche, évalués avec une précision classique. Sur l'ensemble des soumissions, la moyenne est de 0,9822 et la médiane se situe à 0,9868. Les résultats ne présentent pas de différence significative lorsqu'on utilise un test de Student.

Soumission	EDF R&D		Reezocar			Sorbonne			Synapse		
	1	2	1	2	3	1	2	3	1	2	3
Précision	0,9830	0,9868	0,9868	0,9811	0,9849	0,9887	0,9887	0,9887	**0,9906**	0,9849	0,9396

TABLE 6 – Evaluation des prédictions en précision. Le meilleur résultat est en gras

Méthodes Sur cette tâche, des coefficients de similarité ont été employés, tel que le coefficient de Dice par EDF R&D (Cao *et al.*, 2020) ou de plusieurs distances (euclidienne, Manhattan, Minkovski)

par Buscaldi *et al.* (2020) pour l'équipe Sorbonne. Des approches à base de représentations vectorielles plus lourdes ont également été essayées, telle que USE (Universal Sentence Encoder) par Cao *et al.* (2020), ou DRMM (Deep Relevance Matching Model) et MUSE (Multilingual Universal Sentence Encoder) entraînés sur des corpus disponibles en interne et ensuite fournis dans des classifieurs à base de descente stochastique de gradient (SGD) ou de gradient boosting extrême (XGB) par les équipes Reezocar (Tapi Nzali, 2020) et Synapse (Belkacem *et al.*, 2020).

4.3 Tâche 3 – Extraction d'information fine

L'évaluation de la tâche d'extraction d'information est divisée en deux sous-tâches, en fonction de la taille des portions à traiter : une première sous-tâche pour les signes ou symptômes et pathologies (tableau 7), en raison de leur complexité, et une deuxième sous-tâche pour les huit autres catégories (tableau 8). Les résultats ont été calculés au moyen de l'outil BRATeval en évaluation stricte. Sur l'ensemble des soumissions (sauf le premier run de l'équipe Lirmm/Onaos qui ne contenait volontairement aucune prédiction pour cette sous-tâche), la moyenne est de 0,4618 et la médiane se monte à 0,4706 pour la première sous-tâche, tandis que pour la deuxième sous-tâche, la moyenne est de 0,6012 et la médiane s'élève à 0,6151. Notons que, sur ce corpus d'évaluation, les accords inter-annotateur calculés en F-mesure stricte entre les deux annotateurs humains se sont élevés à 0,460 sur les pathologies et 0,470 sur les signes ou symptômes.

Soumission	GLOBAL			Pathologie			Sosy		
	P	R	F	P	R	F	P	R	F
Doing, 1	0,568	0,484	0,523	0,571	0,361	0,443	0,568	0,500	0,532
Doing, 2	0,577	0,493	0,531	0,505	0,337	0,404	0,584	0,513	0,546
Doing, 3	0,532	0,446	0,486	0,434	0,319	0,368	0,543	0,463	0,500
EDF R&D, 1	0,137	0,042	0,065	0,137	0,368	0,199	0,000	0,000	0,000
HESGE, 1	0,576	0,439	0,498	**0,613**	0,295	0,398	0,573	0,458	0,509
HESGE, 2	0,609	0,622	0,615	0,492	0,584	0,534	0,627	0,627	0,627
HESGE, 3	**0,702**	**0,624**	**0,660**	0,575	0,554	0,564	**0,720**	**0,633**	**0,673**
IBM, 1	0,495	0,376	0,427	0,429	0,398	0,413	0,506	0,373	0,430
IBM, 2	0,448	0,419	0,433	0,345	0,416	0,377	0,466	0,419	0,441
Limics, 1	0,422	0,305	0,354	0,264	0,277	0,271	0,454	0,308	0,367
Limics, 2	0,609	0,576	0,592	0,428	**0,645**	0,514	0,649	0,567	0,605
Limics, 3	0,660	0,574	0,614	0,512	0,633	**0,566**	0,689	0,567	0,623
Lirmm/Onaos, 1	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Lirmm/Onaos, 2	0,342	0,294	0,316	0,199	0,416	0,270	0,397	0,278	0,327
Reezocar, 1	0,525	0,372	0,436	0,489	0,386	0,431	0,531	0,371	0,437
Reezocar, 2	0,456	0,331	0,383	0,426	0,259	0,322	0,459	0,340	0,391
Reezocar, 3	0,536	0,397	0,456	0,480	0,440	0,459	0,545	0,391	0,455

TABLE 7 – Evaluation stricte globale et par catégorie (pathologie, signe ou symptôme) en précision, rappel et F-mesure sur la sous-tâche des longues portions. Les meilleurs résultats sont en gras

Méthodes Pour cette tâche, les champs aléatoires conditionnels (CRF de chaîne linéaire) ont majoritairement été employés, notamment par les équipes Doing (Minard *et al.*, 2020) avec des

Soumission	GLOBAL			Anat	Dose	Exam	Mode	Momt	Subs	Trait	Val
	P	R	F	F	F	F	F	F	F	F	F
Doing, 1	0,831	0,606	0,701	0,732	0,347	0,723	0,531	0,726	0,625	0,496	0,802
Doing, 2	**0,839**	0,613	0,708	0,736	0,347	0,731	0,540	0,726	0,643	0,520	0,802
Doing, 3	0,785	0,585	0,670	0,688	0,317	0,701	0,588	0,686	0,580	0,519	0,769
EDF R&D, 1	0,415	0,314	0,358	0,251	0,286	0,465	0,442	0,263	0,432	0,275	0,468
HESGE, 1	0,781	0,507	0,615	0,556	0,368	0,679	0,342	0,627	0,578	0,476	0,797
HESGE, 2	0,737	0,737	0,737	0,798	0,412	0,715	0,639	0,758	0,670	0,557	0,839
HESGE, 3	0,788	0,725	0,755	**0,807**	0,522	0,733	**0,649**	**0,787**	0,638	0,608	**0,856**
IBM, 1	0,743	0,339	0,466	0,152	0,164	0,670	0,527	0,529	0,534	0,510	0,526
IBM, 2	0,695	0,573	0,628	0,662	0,154	0,649	0,587	0,563	0,563	0,438	0,752
Limics, 1	0,659	0,567	0,610	0,694	0,296	0,619	0,453	0,520	0,428	0,406	0,699
Limics, 2	0,764	**0,749**	0,756	0,758	**0,547**	0,802	0,561	0,688	0,769	0,649	0,819
Limics, 3	0,795	0,733	**0,763**	0,763	0,539	**0,805**	0,569	0,701	**0,786**	**0,659**	0,815
Lirmm/Onaos, 1	0,414	0,081	0,135	0,000	0,274	0,000	0,286	0,345	0,000	0,000	0,442
Lirmm/Onaos, 2	0,627	0,508	0,561	0,616	0,245	0,575	0,518	0,000	0,579	0,417	0,671
Reezocar, 1	0,679	0,511	0,583	0,664	0,275	0,470	0,584	0,586	0,599	0,378	0,750
Reezocar, 2	0,681	0,505	0,580	0,653	0,192	0,477	0,561	0,552	0,588	0,380	0,761
Reezocar, 3	0,678	0,530	0,595	0,678	0,260	0,469	0,592	0,633	0,595	0,390	0,768

TABLE 8 – Evaluation stricte globale en précision, rappel et F-mesure, et par catégorie (anatomie, dose, examen, mode, moment, substance, traitement, valeur) en F-mesure uniquement sur la sous-tâche des courtes portions. Les meilleurs résultats sont en gras

modèles CRF pour chaque catégorie, les équipes HESGE (Copara *et al.*, 2020) et Reezocar (Tapi Nzali, 2020). Les modèles de reconnaissance d'entités nommées de l'outil SpaCy fondés sur des réseaux neuronaux convolutifs (CNN) ont également été testés par les équipes Doing et EDF R&D. Les modèles de langue de l'anglais (BERT, BioBERT, RoBERTa), et ceux adaptés au français tels que CamemBERT, dont une version pré-entraînée sur des données biomédicales issues de PubMed, ont également été utilisés par plusieurs équipes dont HESGE et le LIMICS (Wajsbürt *et al.*, 2020) qui a utilisé des versions entraînées sur des corpus multilingues de données web (OSCAR et CCNET) utilisés dans un bi-LSTM avec une couche finale de CRF.

Nous relevons que certaines équipes industrielles ont mis à profit la campagne d'évaluation DEFT pour appliquer sur ces corpus des ressources créées à partir de données métier avec l'outil SpaCy pour l'équipe EDF R&D (Cao *et al.*, 2020), ou d'outils internes tel que WKS (Watson Knowledge Studio) de la suite IBM Watson, algorithme fondé sur l'entropie maximale, pour l'équipe IBM France (Royan *et al.*, 2020). Enfin, l'équipe Lirmm/Onaos (Lemaitre *et al.*, 2020) a utilisé un système à base de règles fondé sur des ressources de la base JeuxDeMots ou du domaine médical. Plusieurs équipes ont témoigné de la difficulté à distinguer les catégories pathologie et signe ou symptôme d'une part, ce que nous reconnaissons pour avoir eut des difficultés à les annoter lors de la préparation des corpus, et à gérer les imbrications d'entités d'autre part.

La complexité des annotations fournies a également poussé l'équipe IBM France a réannoter le corpus en incluant notamment des relations entre entités (localisation, temporalité, précision, etc.) pour exclure les imbrications d'entités. Des règles de post-traitement sont ensuite utilisées pour faire correspondre ces annotations à celles attendues dans la campagne DEFT.

5 Conclusion

L'édition 2020 du défi fouille de texte a proposé trois tâches. Malgré sa complexité, la tâche d'extraction d'informations fine a permis l'obtention des résultats corrects sur les deux catégories d'information les plus difficiles (pathologies et signes ou symptômes), alors même que ces deux catégories ont posé problème lors du travail d'annotation humaine. De plus, les résultats ne dépassaient que rarement les 0,8 de F-mesure sur des catégories a priori plus simples (doses, moments, valeurs). La diversité des contenus dans chacune des catégories semble expliquer ces résultats.

La tâche d'identification de la phrase la plus similaire d'une phrase source parmi trois cibles fournies a permis à l'ensemble des participants d'obtenir d'excellents résultats, qui varient de 0,94 à 0,99 de précision. Le corpus fourni pour cette deuxième tâche semble manifestement facile à traiter de manière automatique. Par contre, lorsqu'il s'agit d'attribuer un degré de similarité sur une échelle à six valeurs, la tâche paraît plus complexe. L'absence volontaire de définition du contenu de chaque degré de similarité et le nombre relativement élevé de degrés disponibles (six degrés de 0 à 5) réduisent les chances de succès.

Références

BELKACEM T., TEISSEDRE C. & ARENS M. (2020). Similarité Sémantique entre Phrases : Apprentissage par Transfert Interlingue. In *Actes de DEFT*, Nancy, France.

BUSCALDI D., FELHI G., GHOUL D., LE ROUX J., LEJEUNE G. & ZHANG X. (2020). Calcul de similarité entre phrases : quelles mesures et quels descripteurs ? In *Actes de DEFT*, Nancy, France.

CAO D., BENAMAR A., BOUMGHAR M., BOTHUA M., OULD-OUALI L. & SUIGNARD P. (2020). Participation d'EDF R&D à DEFT 2020. In *Actes de DEFT*, Nancy, France.

CARDON R. & GRABAR N. (2020). A French corpus for semantic similarity. In *LREC 2020*, p. 1–12.

COPARA J., KNAFOU J., MORO C., RUCH P. & TEODORO D. (2020). Contextualized French Language Models for Biomedical Named Entity Recognition. In *Actes de DEFT*, Nancy, France.

DRAME K., SAMBE G., DIOP I. & FATY L. (2020). Approche supervisée de calcul de similarité sémantique entre paires de phrases. In *Actes de DEFT*, Nancy, France.

GRABAR N. & CARDON R. (2018). CLEAR-Simple Corpus for Medical French. In *Proc of ATA*, Tilburg, The Netherlands. HAL : halshs-01968355.

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). CAS : French Corpus with Clinical Cases. In *Proc of LOUHI*, p. 122–128, Brussels, Belgium. DOI : 10.18653/v1/W18-5614.

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*, Toulouse, France. HAL : hal-02280852.

GROUIN C., GRABAR N., HAMON T. & CLAVEAU V. (2019). Clinical Case Reports for NLP. In *Proc of BioNLP*, Florence, Italy. DOI : 10.18653/v1/W19-5029.

LEMAITRE T., GOSSET C., LAFOURCADE M., PATEL N. & MAYORAL G. (2020). DEFT 2020 – Extraction d'information fine dans les données cliniques : terminologies spécialisées et graphes de connaissance. In *Actes de DEFT*, Nancy, France.

MINARD A.-L., ROQUES A., HIOT N., ALVES M. H. F. & SAVARY A. (2020). DOING@DEFT : cascade de CRF pour l'annotation d'entités cliniques imbriquées. In *Actes de DEFT*, Nancy, France.

ROYAN C., LANGÉ J.-M. & ABIDI Z. (2020). Extraction d'information de cas cliniques avec un système commercial générique. In *Actes de DEFT*, Nancy, France.

STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). brat : a Web-based Tool for NLP-Assisted Text Annotation. In *Proc of EACL*, p. 102–107, Avignon, France.

TAPI NZALI M. (2020). DEFT 2020 : détection de similarité entre phrases et extraction d'information. In *Actes de DEFT*, Nancy, France.

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. DOI : 10.1136/amiajnl-2011-000203.

WAJSBÜRT P., TAILLÉ Y., LAINÉ G. & TANNIER X. (2020). Participation de l'équipe du LIMICS à DEFT 2020. In *Actes de DEFT*, Nancy, France.