Extraction d'information de spécialité avec un système commercial générique

Clothilde Royan, Jean-Marc Langé, Zied Abidi

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

Extraction d'information de spécialité avec un système commercial générique

Clothilde Royan[1,2], Jean-Marc Langé [2], Zied Abidi [2]
(1) Université de Paris 6, Master Ingéniérie des Systèmes Intelligents
(2) IBM France, 17 Avenue de l'Europe, 92275 Bois-Colombes
clothilde.Royan@ibm.com, jmlange@fr.ibm.com,
zied.abidi@fr.ibm.com

RÉSUMÉ

Nous avons participé à la tâche 3 du Défi Fouille de texte 2020, dédiée à l'extraction d'information de spécialité, dans le but de tester notre produit commercial d'extraction d'information, Watson Knowledge Studio (WKS), face à des équipes académiques et industrielles.

Outre la quantité réduite de données d'apprentissage, la nature des annotations des corpus de référence posait des problèmes d'adaptation à notre produit. Aussi avons-nous dû modifier le schéma d'annotation du corpus d'apprentissage, exécuter l'apprentissage, puis appliquer des règles aux résultats obtenus afin d'obtenir des annotations conformes au schéma initial.

Nous avons également appliqué des dictionnaires de spécialité (anatomie, pathologie, etc.) pour injecter de la connaissance du domaine et renforcer les modèles d'apprentissage automatique.

Au final, nos résultats lors de la phase de test se situent dans la moyenne de l'ensemble des équipes, avec des F-mesures de 0,43 pour la sous-tâche 1 et 0,63 pour la sous-tâche 2.

ABSTRACT

Extracting Medical Information with an Off-the-shelf Software Product

We participated in the DEFT 2020 challenge, task 3, to benchmark our software product IBM Watson Knowledge Studio against academic and industry teams, in a demanding information extraction task based on clinical reports.

The data and annotation scheme was challenging for our software, so we change the original DEFT annotation scheme in order to simplify it to avoid embedded annotations and lengthy annotation spans. We apply rules to recombine the results from the ML model into annotations conformant with the original scheme.

We also use medical dictionaries to boost the ML models.

Our final results are very close to the mean values of all participating teams: F1=0,43 on subtask 1, F1=0,63 on subtask 2.

Atelier DÉfi Fouille de Textes, pages 79–90.
HAL : hal-02784744.

MOTS-CLÉS : extraction d'information, données cliniques, Watson Knowledge Studio
KEYWORDS : information extraction, clinical data, Watson Knowledge Studio

1 Introduction

La gamme IBM Watson propose un ensemble de services prêts à l'emploi pour différentes applications de traitement du langage ou de l'image basées sur l'intelligence artificielle : extraction d'information textuelle générale ou spécialisée, recherche sémantique, reconnaissance et synthèse vocale, reconnaissance d'image. Le Défi Fouille de texte 2020 (Cardon et al., 2020), et en particulier la tâche 3, était une bonne occasion de tester notre produit commercial d'extraction d'information, Watson Knowledge Studio (WKS), dans une compétition transparente entre des équipes académiques et industrielles.

2 Description des données

Les données d'entraînement pour la tâche 3 sont un ensemble de 100 cas cliniques ; pour chaque cas, on dispose d'un document contenant le texte (.txt), et d'un document au format Brat (.ann) contenant les annotations identifiées dans ce texte.

Les documents sont de longueur variable, entre 76 et 1407 mots (moyenne 361), avec une majorité de documents relativement courts (autour de 300 mots). Les textes sont spécialisés dans le domaine médical (comme prévu), et plus particulièrement en urologie, avec un vocabulaire très spécialisé.

Les données d'annotation concernent différents types d'information pour les besoins de la tâche 3 : signe ou symptôme (abrégé sosy) et pathologie pour la sous-tâche 1, anatomie, dose, examen, mode, moment, substance, traitement, valeur pour la sous-tâche 2. Dans la suite de ce document, nous mentionnerons ces catégories en majuscules (e.g. PATHOLOGIE). Dans le corpus d'entraînement, les catégories sont diversement représentées, entre 1831 instances de SOSY et 243 instances de MODE.

Les annotations de la sous-tâche 1, catégories SOSY et PATHOLOGIE, sont réputées constituer des « portions assez vastes » (DEFT 2020, 2020), qu'en est-il exactement ? La Figure 3 montre que les SOSY comprennent un nombre significatif d'instances de plus de 5 mots. Cela aura son importance lors du choix de la méthodologie à appliquer, au regard des bonnes pratiques recommandées dans l'utilisation de l'outillage Watson Knowledge Studio.

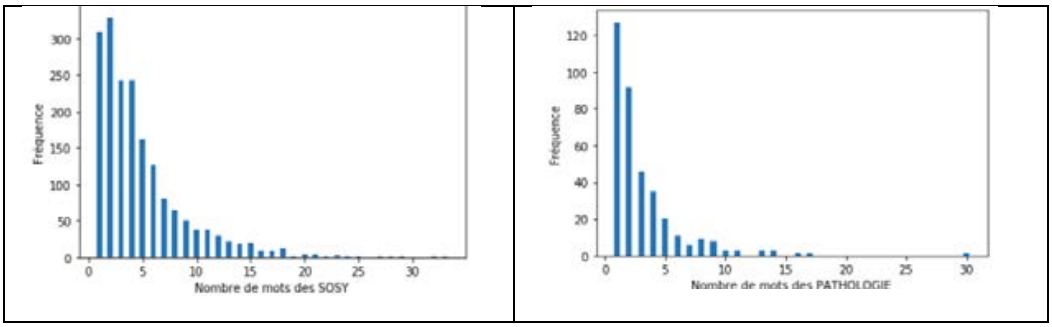

Figure 1 : portée (en mots) des annotations SOSY et PATHOLOGIE

Nous avons analysé un certain nombre d'instances de SOSY, notamment les instances de longue portée. La question qui se posait était alors : comment fait-on pour détecter une annotation d'une portée de plusieurs dizaines de mots ? Il semble facile de repérer où elles commencent, mais détecter où elles s'arrêtent l'est beaucoup moins. Le guide d'annotation (DEFT 2020b) conseille à ce sujet : *Les frontières de ces portions doivent être envisagées sous l'angle d'une seule idée par portion annotée, même si plusieurs informations peuvent venir compléter cette idée.*

3 Méthode

3.1 Aperçu Général

Nous présentons dans cette section l'outillage d'apprentissage et la méthode et l'outillage utilisés pour les deux sous-tâches ; nous aborderons les spécificités des sous-tâches dans les sections suivantes.

3.1.1 Watson Knowledge Studio

L'outil privilégié pour les tâches d'extraction d'information dans le portefeuille IBM d'IA est Watson Knowledge Studio (WKS). Nous avons choisi de participer à la campagne DEFT 2020 avec pour objectif de mesurer la performance de ce système et de le comparer à la « concurrence » sur une tâche *a priori* ardue.

WKS fait partie de la gamme IBM Watson, qui propose un ensemble de services d'IA disponibles dans le cloud. La philosophie de ces services est de fournir une IA prête à l'emploi : un service immédiatement utilisable, libérant les utilisateurs des préoccupations de choix d'algorithme, réglage des paramètres, etc. À la différence de la plupart des services Watson, disponibles sous forme d'interface de programmation (API), WKS propose quant à lui, dans une interface Web unique, de gérer le processus complet de fabrication de modèles d'extraction de connaissances à base d'apprentissage automatique :

- création du modèle de données,
- collecte des documents,
- gestion des corpus d'apprentissage/test
- préannotation automatique et annotation humaine du corpus,

- gestion des équipes d'annotateurs et de l'accord inter-annotateurs,
- préannotation automatique du corpus (avec modèles existants, dictionnaires ou règles)
- création du modèle d'apprentissage
- gestion des tests avec tableaux de bord des résultats
- versionnage des modèles d'apprentissage
- déploiement des modèles d'apprentissage

Le modèle de données (*type system*) est un modèle classique d'entités et relations. WKS permet également de gérer les *coréférences*, liens permettant de lier les mentions d'une même entité, comme pour la reprise anaphorique par des pronoms.

Il est possible -optionnellement- de renforcer l'apprentissage des modèles au moyen de dictionnaires qui contribuent à l'apprentissage sans pour autant oblitérer l'influence des autres attributs calculés à partir du contexte textuel. WKS propose en outre un **moteur de règles** permettant, dans une interface visuelle, de concevoir des modèles d'extraction de connaissances s'appuyant sur des dictionnaires et des règles.

D'un point de vue technique, le module d'apprentissage de WKS s'appuie sur un moteur de classification de séquences basé sur l'entropie maximale, descendant du système décrit dans (Radu et al., 2004).

3.1.2 WKS, adapté au Défi ?

Le fait de proposer une démarche de bout en bout, depuis l'annotation humaine jusqu'au déploiement des modèles, impose un certain nombre de contraintes qui justifient la nécessité d'adapter notre démarche en regard des spécificités des données proposées dans DEFT 2020.

En effet, WKS prescrit en partie son modèle d'annotation ; ainsi la documentation du produit (IBM, 2019) préconise les bonnes pratiques suivantes pour l'annotation d'entités :

- annoter des passages plutôt **courts** (de préférence sur 1 ou 2 mots)
- **éviter** absolument les **imbrications** d'entités. Il est préconisé, pour ce faire, d'utiliser des relations entre deux entités pour former des instances de concepts d'une portée plus étendue.

Nous savions dès le départ, avec la définition de la tâche 3 donnée en page principale du Défi 2020, que nous allions devoir nous adapter pour résoudre cette tâche.

3.1.3 Evolution de la démarche

3.1.3.1 Démarche, premier temps : utilisation du corpus d'apprentissage non modifié

Pour produire une ligne de base, nous avons dans un premier temps écrit un script de conversion du corpus d'apprentissage au format BRAT vers le format d'ingestion de WKS. Les phases classiques de tokenisation et segmentation en phrases sont effectuées

automatiquement par WKS lors de l'ingestion. La segmentation en phrases n'a posé que quelques rares problèmes, où des expressions chiffrées contenant un point sont coupées. Du fait de la rareté de ce phénomène, nous n'avons pas cherché à le rectifier.

Les figures 4 et 5 montrent un extrait du corpus d'apprentissage au format Brat et au format WKS :

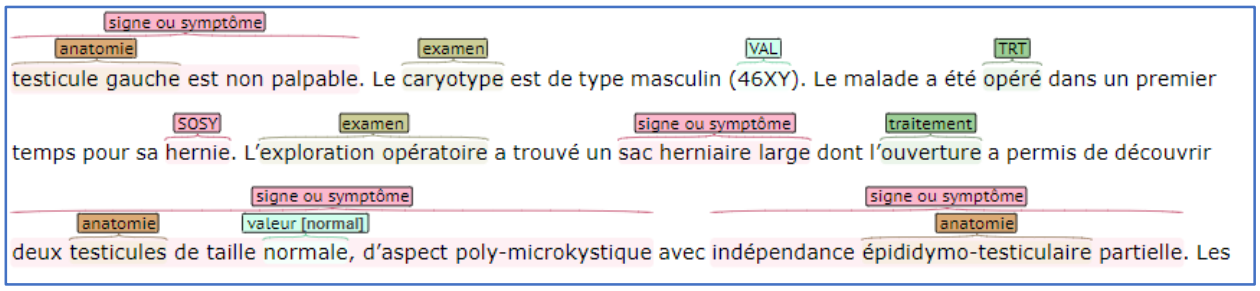

Figure 2 : extrait du corpus affiché dans BRAT

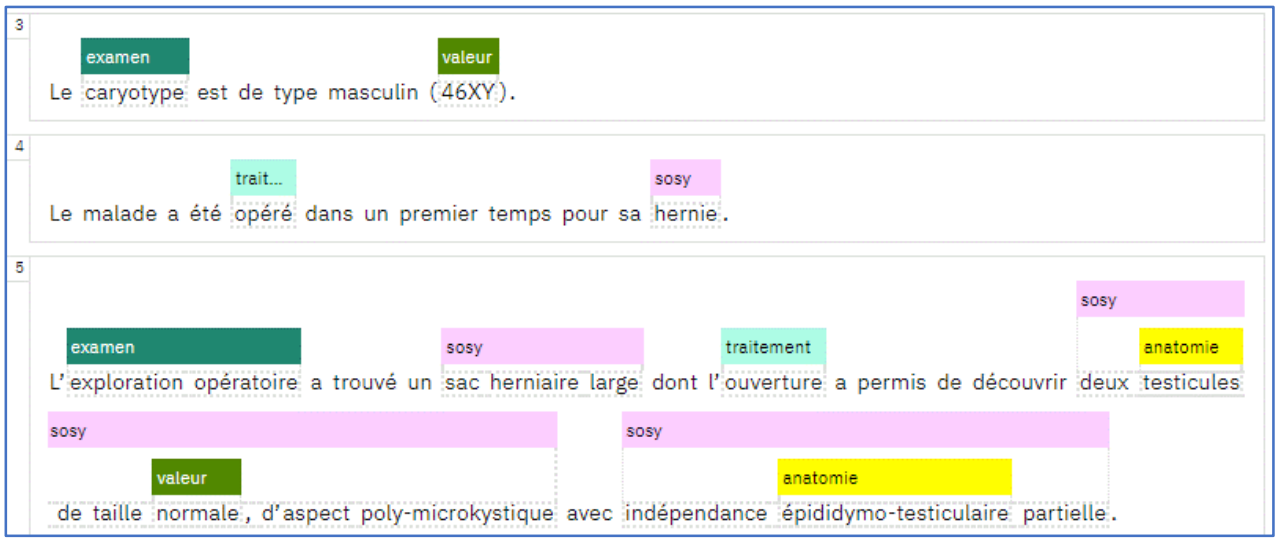

Figure 3 : extrait du corpus affiché dans WKS

Une fois le corpus disponible en format WKS, nous avons produit un premier modèle WKS, que nous avons testé sur une portion du corpus d'entraînement non vue durant l'entraînement.

Les résultats de tests avec cette approche « corpus brut » sont très faibles. Voici par exemple le résultat du modèle entraîné sur la totalité (100 documents) du corpus d'entraînement et testé sur le corpus de test DEFT 2020 :

	Précision	Rappel	F1
pathologie	0.03	0.31	0.05
sosy	0.03	0.29	0.06
anatomie	0.39	0.12	0.18
dose	0.19	0.33	0.24
examen	0.13	0.63	0.22
mode	0.00	0.00	0.00

moment	0.25	0.56	0.35
substance	0.12	0.42	0.18
traitement	0.07	0.47	0.11
valeur	0.10	0.53	0.17

Figure 4 : résultats du modèle "brut"

Pour tenter d'expliquer ce résultat, comparons un passage annoté du corpus de test avec la prédiction du modèle sur ce même passage :

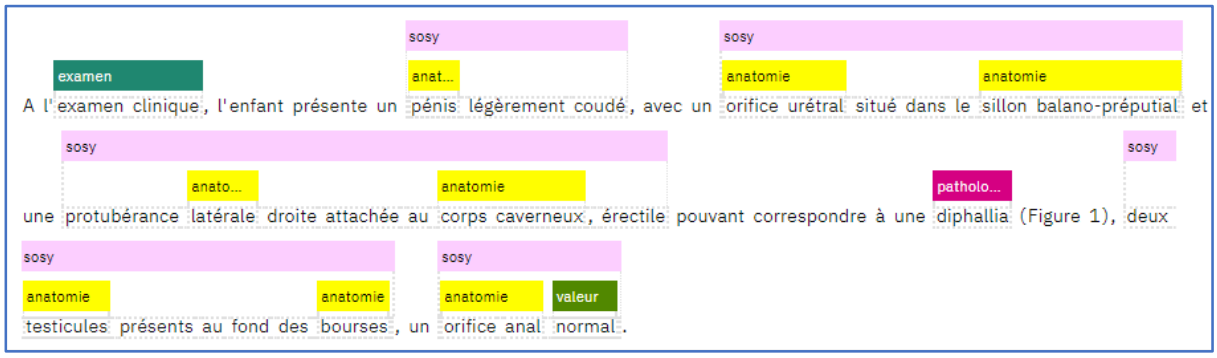

Figure 5 : Annotations du corpus de référence (interface Watson Knowledge Studio)

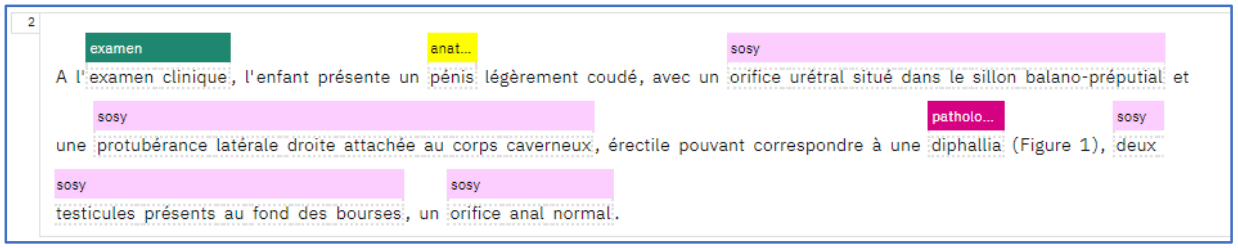

Figure 6 : Annotations prédites par le modèle "brut" (interface Watson Knowledge Studio)

Le problème est très clair : lorsque des annotations sont imbriquées (empilées dans la figure 7 ci-dessus), le modèle WKS ne prédit que la catégorie englobante (tous les SOSY sauf le premier), ou bien la ou les catégories englobées (e.g. *pénis* comme ANATOMIE dans la première phrase). De plus, les annotations longues semblent systématiquement oblitérer les annotations englobées, alors que dans le cas des annotations courtes, ce sont les englobées qui survivent. Or il y a beaucoup plus d'annotations courtes que longues dans le corpus, même pour les SOSY ; voilà qui explique les mauvais scores. Les bonnes pratiques de la documentation de WKS semblent pertinentes.

Nous avons donc dû adapter notre méthode aux contraintes posées par WKS, à savoir non-imbrication et courte portée des annotations. Nous avons donc défini un nouveau processus de traitement avec trois grandes étapes :

- **réannotation** du corpus dans WKS pour dégager des annotations plus courtes (entités au sens WKS), et en remplaçant les imbrications par des relations : cette étape a été en partie automatisée pour certaines imbrications, mais l'essentiel du travail reposait sur des annotateurs humains dans l'interface de WKS ;
- **création d'un modèle** basé sur ces réannotations ;

- **recombinaison** des entités et relations prédites par ce modèle, afin de retrouver des annotations conformes au corpus d'apprentissage DEFT 2020, notamment pour obtenir toute la portée des annotations longues telles que SOSY et PATHOLOGIE ; pour ce faire, nous avons conçu des règles de recombinaison.

Nous détaillons ces différentes étapes dans la section suivante.

3.1.3.2 Démarche, deuxième temps : réannotation du corpus et recomposition

Pour mettre en œuvre la réannotation du corpus, il nous faut au préalable examiner la structure des imbrications d'annotation pour dégager le jeu d'entités et relations, et les pratiques d'annotation qui permettent de restituer au mieux, *in fine*, les annotations originelles.

Etude des annotations longues/imbriquées

Pour étudier les patrons d'imbrication d'annotations, en particulier des annotations longues, afin de dégager le schéma d'annotation alternatif à utiliser dans WKS, nous avons procédé comme suit :

1. récupération de toutes les annotations imbriquées ;
2. suppression des mots-outils ;
3. extraction des mots de tête des annotations « coiffantes » ;
4. remplacement des passages de texte par la catégorie d'annotation ;
5. stockage du vocabulaire des passages de texte restant hors des annotations et remplacement par une marque « MOT ».

Une fois identifiés ces patrons, nous pouvons les visualiser avec des couleurs affectées à certaines entités. En parallèle, nous avons extrait le vocabulaire des passages annotés et des passages inter-annotations avec les fréquences des mots.

Cela nous a permis d'identifier les patrons les plus saillants de combinaisons d'annotation. À titre d'exemple, voici un exemple de visualisation des motifs de SOSY, ainsi qu'un extrait du vocabulaire utilisé dans ANATOMIE :

examen MOT **valeur** (77) **examen** **valeur** (56) **anatomie** MOT (17) **examen** : **valeur** 10)	gauche (31) droit (18) droite (17) rein (13) inférieure (6)	supérieure (6) inférieur (4) supérieur (4) vessie (4) col (4)

Figure 7: Patrons de SOSY et vocabulaire d'ANATOMIE

Cela nous indique clairement le genre de de réannotation que nous devons mener à bien, et les règles de recomposition à appliquer pour obtenir l'annotation finale. Cela nous suggère aussi l'opportunité de créer une nouvelle catégorie, POSITION, (*gauche*, *droit*, *inférieur*, *supérieur*…) pour mieux annoter les ANATOMIE.

Réannotation

Suite à l'étude des patrons d'imbrication, nous avons conclu à la nécessité d'étendre le jeu de catégories DEFT 2020 en y ajoutant :

- deux catégories :
  - POSITION pour rendre compte que les ANATOMIE sont fréquemment composées d'un concept d'anatomie proprement dit et de précisions de localisation : « **raphé** *antérieur* », « *extrémité proximale* du **moignon urétéral** *gauche* »
  - CARACTERE pour rendre compte de différentes précisions affectées aux têtes de SOSY et autres : **obésité** *importante*, **rétention vésicale** *complète*
- des relations entre ces catégories, afin de couvrir des portées de texte se rapprochant de l'annotation originale :

Relation	Entité source	Entité cible
A_LIEU	SOSY, PATHOLOGIE, EXAMEN, TRAITEMENT	ANATOMIE
A_POSITION	ANATOMIE et autres	POSITION
A_VALEUR	EXAMEN, SUBSTANCE	VALEUR
PORTE_SUR	TRAITEMENT, SOSY, autres	PATHOLOGIE, SUBSTANCE, …
A_CARACTÈRE	SOSY, PATHOLOGIE, EXAMEN, TRAITEMENT	CARACTÈRE
A_PRECISIONS	EXAMEN, TRAITEMENT	EXAMEN, TRAITEMENT
A_TEMPORALITE	SOSY, PATHOLOGIE, EXAMEN, TRAITEMENT	MOMENT

Figure 8 : Schéma d'annotation modifié

Nous avons ensuite procédé à modifier le corpus en appliquant ce nouveau schéma d'annotations, sans imbrication d'entités, et en visant des annotations les moins longues possibles grâce au nouveau modèle d'entités-relations. Le travail a été partagé par deux annotateurs humains, sans contrôle inter-annotateurs par manque de temps. La figure suivante montre un extrait d'un document réannoté :

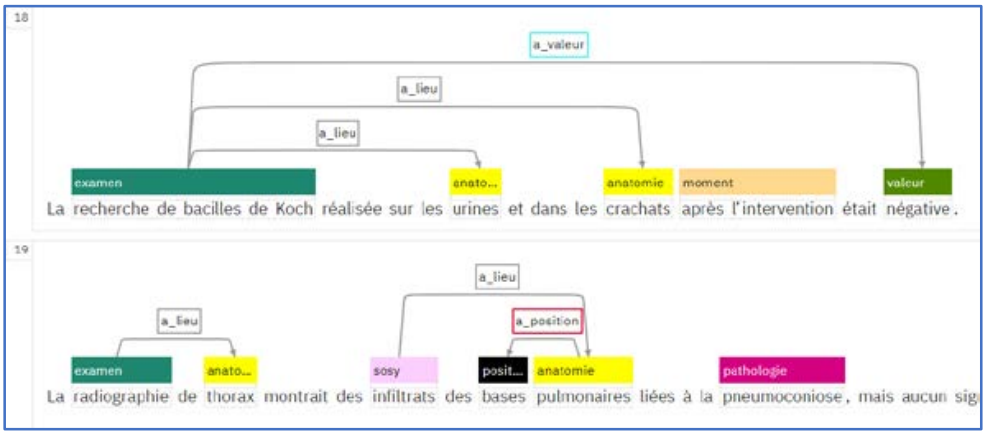

Figure 9 : Exemple de réannotation

Règles de recombinaison des entités-relations

Une fois le corpus d'entraînement réannoté, nous pouvons produire des modèles, mais les résultats de ces modèles utilisent notre schéma d'annotations modifié. Il nous faut donc recombiner, au moyen de règles, les entités et relations obtenues, afin de retrouver

des annotations conformes au corpus d'apprentissage Golden DEFT 2020, notamment pour obtenir toute la portée des annotations longues telles que SOSY et PATHOLOGIE.

Après application du modèle WKS sur le corpus de test, nous convertissons le résultat au format BRAT (.ann), puis nous appliquons les règles comme suit :

1. Reconstruction des annotations le long des relations A_CARACTERE et A_POSITION : les entités CARACTERE et POSITION constituent une brique de base dans plusieurs annotations, notamment SOSY, et il faut donc les réincorporer aux annotations de tête, auxquelles elles sont liées par les relations.

2. Reconstruction des autres relations (sauf EXAMEN <A_VALEUR> VALEUR qui fait l'objet d'une règle particulière, cf plus bas): les différentes relations qui lient les entités annotées permettent de reconstituer les annotations longues. Une fois les deux entités trouvées, nous fusionnons la mention en prenant les bornes les plus larges puis actualisons l'entité correspondant à la tête de l'annotation avec les nouvelles valeurs, avant de supprimer l'entité correspondant à la fin de l'annotation.

3. Pour les relations de type « EXAMEN <A_VALEUR> VALEUR », ce type de relation doit être annoté systématiquement comme SOSY, tout en conservant les entités EXAMEN et VALEUR dans les annotations. Nous lions d'abord des entités voisines (1 avant et 1 après) puis explorons un contexte jusqu'à 20 entités avant et 20 après l'entité qui nous intéresse. Finalement, nous cherchons des ponctuations (point, virgule, saut de ligne) qui invaliderait la qualité d'annotation.

4. Reconstruction « FREQUENCE – X – FREQUENCE » : la visualisation des patrons de combinaison d'annotations (cf supra) nous a appris que le patron « FREQUENCE – X – FREQUENCE » revenait souvent (par ex. « *4 épisodes* de vomissement *par jour* »). Dans le corpus d'entraînement, cette combinaison forme une seule annotation FREQUENCE. Pour reconstruire l'annotation canonique, nous cherchons les entités FREQUENCE puis dans leur voisinage (à 2 entités près) une autre entité FREQUENCE. Nous vérifions que la portée résultant de la fusion ne comporte pas de ponctuation ou de saut de ligne qui invaliderait le patron.

5. Reconstruction de la juxtaposition des SOSY : pour reconstruire des SOSY longs qui seraient une juxtaposition de SOSY plus courts (par exemple : *une masse tumorale, de 6 cm de diamètre avec une forme* …), nous vérifions si la position de début d'une annotation est distante de moins de *n* caractères de la valeur de fin d'une autre annotation. Si tel est le cas, nous fusionnons les annotations en prenant les bornes les plus larges puis actualisons ces valeurs dans l'entité correspondant à la tête de l'annotation.

6. Vérification des imbrications (éviter SOSY dupliqués) : Nous avons remarqué que certaines annotations étaient présentes deux fois, une fois en tant que « petite » annotation et une autre fois au sein d'une « grande » annotation. Nous vérifions toutes les annotations pour éviter cette incohérence.

7. Nettoyage des entités CARACTERE et POSITION sans relation : nous supprimons ces entités si elles sont orphelines.

Résultats de l'approche « réannotation »

La version la plus aboutie du système de règles de recomposition nous permet d'atteindre les scores suivants (ces scores sont les valeurs finales pour la tâche et tiennent compte de l'amélioration du modèle WKS avec l'apport de dictionnaires que nous verrons plus loin) :

3.1	Précision	Rappel	F1
pathologie	0.35	0.42	0.38
sosy	0.47	0.42	0.44
Overall	**0.45**	**0.42**	**0.43**

3.2	Précision	Rappel	F1
anatomie	0.75	0.59	0.66
dose	0.23	0.12	0.15
examen	0.66	0.64	0.65
mode	0.72	0.49	0.59
moment	0.72	0.46	0.56
substance	0.60	0.53	0.56
traitement	0.52	0.38	0.44
valeur	0.83	0.69	0.75
Overall	**0.70**	**0.57**	**0.63**

Figure 10 : Évaluation après application des règles de recomposition

Cette approche permet donc d'améliorer nettement les résultats par rapport à l'approche « corpus brut », mais elle reste à l'évidence limitée par les erreurs dans la détection des « petites » annotations et par les imperfections du jeu de règles. Nous n'avons pas eu suffisamment de ressources pour explorer plus avant les règles de recombinaison. En revanche, nous avons cherché à améliorer la détection des « petites » entités grâce à l'apport des dictionnaires, ce que nous verrons dans la section suivante.

3.2 Travail spécifique sur la sous-tâche 1

3.2.1 Apport des dictionnaires

WKS offrant la possibilité d'augmenter les attributs du modèle d'apprentissage au moyen de dictionnaires. Lors de l'import d'un dictionnaire, on précise l'entité à laquelle il contribue : en effet, WKS prend en compte les dictionnaires comme un attribut de plus dans la modélisation du contexte d'une entité, sans leur donner la priorité.

Nous avons utilisé des dictionnaires pour « fortifier » les annotations SOSY et PATHOLOGIE, ainsi que celles qui sont souvent imbriquées dans ces dernières, telles qu'ANATOMIE, POSITION (*supérieur, antérieur, basal, apical*…), ou EXAMEN. Ces dictionnaires sont constitués par conversion de ressources terminologiques fournis par les classifications médicales internationales telles que SNOMED ou CIM10, ou récupérés de façon ad hoc sur certains sites Web ou Wikipedia.

Le travail de collecte, nettoyage et enrichissement des dictionnaires (collecte, triage, lemmatisation, génération des formes alternatives, vérification de non-recouvrement

entre deux dictionnaires) est une tâche très chronophage et nous n'avons pas pu constituer des dictionnaires très complets et très propres dans le temps imparti, mais ils contribuent néanmoins à améliorer les résultats. À titre d'exemple, la figure ci-dessous indique l'apport (en différence de précision, rappel et F-mesure par rapport à un modèle calculé sans dictionnaire, appliqué au corpus de test), d'un dictionnaire de symptômes lié aux SOSY (à gauche), et d'anatomie, lié à ANATOMIE, à droite :

cat	diff_Pre	diff_Rec	diff_F1
anatomie	0.0047	0.0018	0.0031
dose	0.0192	0.0192	0.0216
examen	0.0035	0.0025	0.0029
mode	0	0	0
moment	0.0026	0.006	0.0052
pathologie	-0.0053	-0.006	-0.0057
sosy	-0.0013	0.0078	0.0039
substance	0.0054	-0.0032	0.0005
traitement	0.0031	-0.0098	-0.0055
valeur	0.0005	0.0023	0.0016
Overall	0.0013	0.0027	0.0021

cat	diff_Pre	diff_Rec	diff_F1
anatomie	0.0453	0.0732	0.0603
dose	-0.01	0	-0.0021
examen	-0.0096	-0.0171	-0.0136
mode	0.0038	0.0112	0.0093
moment	-0.0121	-0.0122	-0.0127
pathologie	-0.0107	-0.006	-0.0088
sosy	0.0047	0.0055	0.0051
substance	-0.0031	0.0032	0.0005
traitement	0.0005	-0.0065	-0.0042
valeur	0	0	0
Overall	0.0105	0.0152	0.0132

Figure 11 : Apport d'un dictionnaire de symptômes d'un dictionnaire d'anatomie

Ces visualisations permettent de jauger l'impact du dictionnaire, minime avec le dictionnaire de symptômes, qui contient des termes de 1 ou 2 mots, alors que les annotations SOSY couvrent typiquement plusieurs mots, voire dizaines de mots. Le dictionnaire d'anatomie a un impact non seulement sur l'entité à laquelle il est lié, mais il affecte également -en mieux ou en pire- les scores d'autres entités. Nos différents tests montrent que l'on peut gagner 3 points de F-mesure avec des dictionnaires rapidement préparés.

3.3 Travail spécifique sur la sous-tâche 2

Nous n'avons consacré que peu de ressources à cette deuxième sous-tâche, mais avons décidé de concourir car nous pouvions produire des résultats par simple apprentissage. Nous avons cependant amélioré, par un travail sur les règles ou par ajout de dictionnaires, les catégories qui contribuent aux SOSY comme EXAMEN, VALEUR, ou ANATOMIE.

4 Résultats et discussion

Au final, nos résultats (F-mesure 0,43 pour la sous-tâche 3.1 et de 0,64 pour la sous-tâche 3.2) sont très proches de la moyenne de l'ensemble des équipes. Cela valide en partie notre approche, qui a consisté à compenser les contraintes liées à notre logiciel en réannotant le corpus d'entraînement avec un schéma d'annotation modifié, puis en appliquant des règles pour recombiner les annotations obtenues afin de reconstituer les annotations d'origine. Le cas d'école est la séparation de ANATOMIE en deux entités (anatomie, position) et une relation entre les deux, qui permet à l'apprentissage automatique, en ciblant mieux ses clients, d'obtenir de meilleurs résultats. Ce principe de réannotation est parfois utilisé -de façon peut-être moins appuyée-, dans nos projets

clients avec WKS, où le travail selon des méthodes agiles induit fréquemment des ajustements du schéma d'annotation pour prendre en compte de nouveaux motifs textuels qui émergent au cours du projet.

Nous avons également amélioré les résultats en utilisant des dictionnaires de domaine.

Nous voyons plusieurs axes d'amélioration:

– Plus de données d'entraînement : notre expérience avec WKS nous indique que l'apport de données supplémentaires permettrait d'atteindre des résultats bien meilleurs pour la sous-tâche 3.2 et partiellement pour la 3.1.
– Automatisation de la réannotation : nous avons réalisé par programmation les réannotations simples, il est parfaitement envisageable de le faire de façon extensive.
– Utilisation d'un vrai moteur de règles : nos règles sont codées, mais cela est difficilement maintenable. WKS dispose d'un moteur de règles qui permettra bientôt d'appliquer les règles sur des informations extraites par le modèle d'apprentissage.
– Analyse syntaxique : extraire des annotations longues reste une tâche ardue, et il faut chercher une autre approche pour la résoudre. Une plus grande familiarité avec le corpus nous indique que les segments longs de SOSY et PATHOLOGIE suivent pour la plupart des motifs grammaticaux classiques : liens de complémentation, subordination et coordination. Les exemples abondent, comme par exemple :

kyste rénal / avec une paroi épaisse / se rehaussant / après injection / de produit de contraste.

Cela suggère l'utilisation d'un analyseur grammatical de dépendance (basé sur des règles ou sur l'apprentissage automatique) pour dérouler la pelote à partir de la tête de l'annotation jusqu'au terme de celle-ci.

Références

CARDON R., GRABAR N., GROUIN C. & HAMON T. (2020). Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques. In: Actes de DEFT

DEFT. (2020a). Défi fouille de texte 2020. https://deft.limsi.fr/2020/

DEFT. (2020b). DEFT 2020: Guide d'annotation. https://deft.limsi.fr/2020/guide-deft.html

FLORIAN R., HASSAN H., ITTYCHERIAH A., JING H., KAMBHATLA N., LUO X., NICOLOV N. & ROUKOS S. (2004). A Statistical Model for Multilingual Entity Detection and Tracking. In *Proceedings of HLT-NAACL 2004*: Boston, Mass., USA

IBM Corp. (2019). Watson Knowledge Studio User's Guide. https://cloud.ibm.com/docs/watson-knowledge-studio?topic=watson-knowledge-studio-user-guide