DÉfi Fouille de Textes 2019: indexation par extraction et appariement textuel

Jean-Christophe Mensonides[1] Pierre-Antoine Jean[1]
Andon Tchechmedjiev[1] Sébastien Harispe[1]
(1) LGI2P, IMT Mines Ales, Univ Montpellier, Ales, France
{prénom}.{nom}@mines-ales.fr

RÉSUMÉ

Cet article présente la contribution de l'équipe du Laboratoire de Génie Informatique et d'Ingénierie de Production (LGI2P) d'IMT Mines Alès au DÉfi Fouille de Textes (DEFT) 2019. Il détaille en particulier deux approches proposées pour les tâches liées à (1) l'indexation et à (2) la similarité de documents. Ces méthodes reposent sur des techniques robustes et éprouvées du domaine de la Recherche d'Information et du Traitement Automatique du Langage Naturel, qui ont été adaptées à la nature spécifique du corpus (biomédical/clinique) et couplées à des mécanismes développés pour répondre aux spécificités des tâches traitées. Pour la tâche 1, nous proposons une méthode d'indexation par extraction appliquée sur une version normalisée du corpus (MAP de 0,48 à l'évaluation) ; les spécificités de la phase de normalisation seront en particulier détaillées. Pour la tâche 2, au-delà de la présentation de l'approche proposée basée sur l'évaluation de similarités sur des représentations de documents (score de 0,91 à l'évaluation), nous proposons une étude comparative de l'impact des choix de la distance et de la manière de représenter les textes sur la performance de l'approche.

ABSTRACT

DEFT 2019 : extraction-based document indexing and textual document similarity matching

This paper presents the contribution of the LGI2P (Laboratoire de Génie Informatique et d'Ingénierie de Production) team from IMT Mines Alès to the DEFT 2019 challenge (DÉfi Fouille de Textes). We detail two approches we devised for the tasks pertaining to (1) the indexing and to (2) the similarity of documents. Said approaches rely on proven and robust techniques from Information Retrieval and Natural Language Processing that have been adapted to the specificities of the corpus (biomedical text) and of the formulation of the tasks. For task 1, we propose an indexing-by-extraction approach applied on the corpus after a normalisation procedure (MAP=0.48) that we will detail further. For task 2, we proposed a similarity-based approach computed on vector representation of the documents (score=0.910) and study the impact of the choice of the similarity metric and of the document representation method on task performance.

MOTS-CLÉS : Indexation de documents, similarité sémantique, recherche d'information, corpus biomédical.

KEYWORDS: Document indexing, semantic similarity, information retrieval, biomedical corpus.

1 Tâche 1 : Indexation des cas cliniques

La tâche d'extraction de mots-clés consiste à distinguer les mots de l'unité linguistique analysée (*e.g.* document, corpus) qui sont caractéristiques de l'unité au regard d'un objectif prédéfini. Elle est généralement exploitée dans l'indexation (Marchand *et al.*, 2016) et le résumé de documents textuels (Gupta & Lehal, 2010). Les mots-clés peuvent dans certains cas être conceptualisés comme un sous-ensemble de méta-données de nature sémantique associées à ces documents. La stratégie utilisée pour l'obtention des mots-clés permet de distinguer deux types d'approches au sein de la tâche d'indexation de documents : les approches par extraction et les approches par assignation (Chartier *et al.*, 2016). La principale différence entre ces deux types d'approches repose sur la provenance des mots-clés. Les approches par extraction s'emploient à annoter un document avec des mots issus du document, tandis que les approches par assignation visent à aligner un document avec une liste de mots-clés issus d'un vocabulaire contrôlé sans nécessairement contraindre les mots-clés sélectionnés à apparaître dans le document.

Le premier type d'approche, par extraction, repose principalement sur une évaluation de la *pertinence* des termes d'un document pour sa caractérisation. La notion de pertinence peut être abordée de différentes manières, *e.g.* statistique, par apprentissage supervisé ou non-supervisé. L'étude statistique permet notamment d'évaluer les termes au travers de leur usage au sein des documents et du corpus. Une des approches les plus courantes repose sur le calcul du coefficient de pondération TF-IDF (Jones, 1972). Ce coefficient cherche à retranscrire l'importance relative d'un terme par rapport à un document et à l'ensemble du corpus. Sur la base de ce modèle, l'importance d'un terme sera d'autant plus grande si le terme apparaît fréquemment dans le document et peu fréquemment dans le corpus. D'autres exemples de coefficients de pondération employés dans le domaine de la classification de textes pour la pondération de termes conditionnée à une classe donnée peuvent également être cités *e.g.* le RDF (*Relevant Document Frequency*), l'IG (*Information Gain*), le MI (*Mutual Information*) ou bien encore le *Chi Square* (Nanas *et al.*, 2003; Hamdan, 2015).

La pertinence d'un terme pour caractériser un document peut également être évaluée à partir de méthodes d'apprentissage supervisé et non-supervisé. Parmi les méthodes d'apprentissage supervisé, KEA - *Keyphrase Extraction Algorithm* (Witten *et al.*, 2005) - exploite un modèle bayésien en utilisant le TF-IDF et le ratio de la première occurrence du mot-clé dans un texte comme fonctions caractéristiques. Une autre approche, proposée par Zhang (2008), se base sur un CRF (*Conditional Random Field*) qui exploite des fonctions caractéristiques liées à la position (*e.g.* présence/absence d'un mot au sein du résumé ou dans le corpus du texte), lexicales (*e.g.* fonction grammaticale du mot) et statistiques (*e.g.* TF-IDF) afin d'identifier et de généraliser les mots-clés au sein des textes. Concernant les modèles d'apprentissage non supervisé, le LSI (*Latent Semantic Indexing*) est un exemple d'approches fréquemment utilisées (Deerwester *et al.*, 1990). Elle se base sur une décomposition matricielle de type SVD (*Singular Value Decomposition*) appliquée à une matrice termes-documents décrivant les occurrences des termes dans les documents. Cette approche permet alors d'indexer les documents par un ensemble de *concepts* (compositions linéaires des représentations des termes dans la matrice initiale).

Le second type d'approche, par abstraction, se focalise plus particulièrement sur le rapprochement sémantique des termes. Ce type de procédé peut s'appuyer sur des mesures de similarité sémantique basées sur une structuration de la connaissance fournie *a priori* et/ou sur l'analyse de corpus de textes. Les mesures de similarité sémantique basées sur une structuration de la connaissance (*e.g.* ontologie, vocabulaire structuré) permettent d'évaluer la similarité de sens entre concepts/termes

selon l'information modélisée par la structuration de la connaissance (Harispe *et al.*, 2014). Par exemple, l'approche USI - *User-oriented Semantic Indexer* (Fiorini *et al.*, 2015) - exploite cette stratégie d'abstraction afin d'indexer des articles biomédicaux.

Les approches basées sur l'analyse de corpus de textes reposent très souvent explicitement ou implicitement sur l'hypothèse distributionnelle qui considère que le sens d'un terme est donné par son voisinage dans les textes, *i.e.* son usage. On distingue aujourd'hui les analyses traditionnelles des fréquences de co-occurrences, des approches par plongements sémantiques de mots récemment popularisées par l'approche WORD2VEC (Mikolov *et al.*, 2013). Des travaux liés au résumé textuel par abstraction démontrent l'efficacité d'employer ce type d'approche couplée à un réseau neuronal récurrent (Nallapati *et al.*, 2016).

Enfin, nous pouvons évoquer des approches hybrides dont la spécificité reposent sur la manière de représenter les textes sous la forme de graphes afin d'en extraire des mots-clés pertinents par extraction ou par abstraction. Ces approches s'appuient sur des algorithmes de parcours de graphes pondérés (Wang *et al.*, 2014; Mahata *et al.*, 2018) ou non pondérés (Litvak & Last, 2008) ou bien de recherche de motifs *e.g.*, recherche des cliques maximales (Kim *et al.*, 2014). Ces graphes peuvent être dirigés dans le cas par exemple où le graphe est construit en tenant compte de la succession des termes ou non dirigés si le graphe s'appuie sur une matrice de co-occurrences.

La stratégie d'indexation définie par notre équipe pour cette première tâche a été déterminée suite à l'étude statistique du corpus présentée en section 1.1. La section 1.2 présente la méthodologie mise en place et les différents résultats obtenus.

1.1 Description du corpus et métriques d'évaluation

1.1.1 Description et statistiques descriptives du corpus

Les tâches de cette quinzième édition du défi DEFT reposent sur un corpus d'entraînement constitué de 290 couples de textes de cas clinique/discussion (Grabar *et al.*, 2019). A chaque couple est associé un ensemble de mots-clés qui ont été définis manuellement au terme d'un consensus entre deux annotateurs. L'ensemble des mots-clés utilisés pour annoter les couples cas clinique/discussion forment un vocabulaire contrôlé. Le tableau 1 résume différentes statistiques liées aux données fournies dans le cadre du défi DEFT 2019.

# couples cas clinique/discussion dans $\mathcal{C}$	290
# moyen de mots dans les cas cliniques de $\mathcal{C}$	332
# moyen de mots dans les discussions de $\mathcal{C}$	764
# de mots-clés dans le vocabulaire contrôlé	1311
# de mots-clés utilisés dans $\mathcal{C}$	1123 (0.85%)
# de mots-clés avec une correspondance exacte dans les cas cliniques de $\mathcal{C}$	441
# de mots-clés avec une correspondance exacte dans les discussions de $\mathcal{C}$	658
# de mots-clés abstraits au sein d'un couple de $\mathcal{C}$	390

TABLE 1 – Statistiques sur les couples cas clinique/discussion et sur les mots-clés associés. Avec $\mathcal{C}$ le jeu d'entraînement et # un symbole signifiant nombre.

La figure 1 présente également des statistiques intéressantes liées au corpus et au vocabulaire contrôlé.

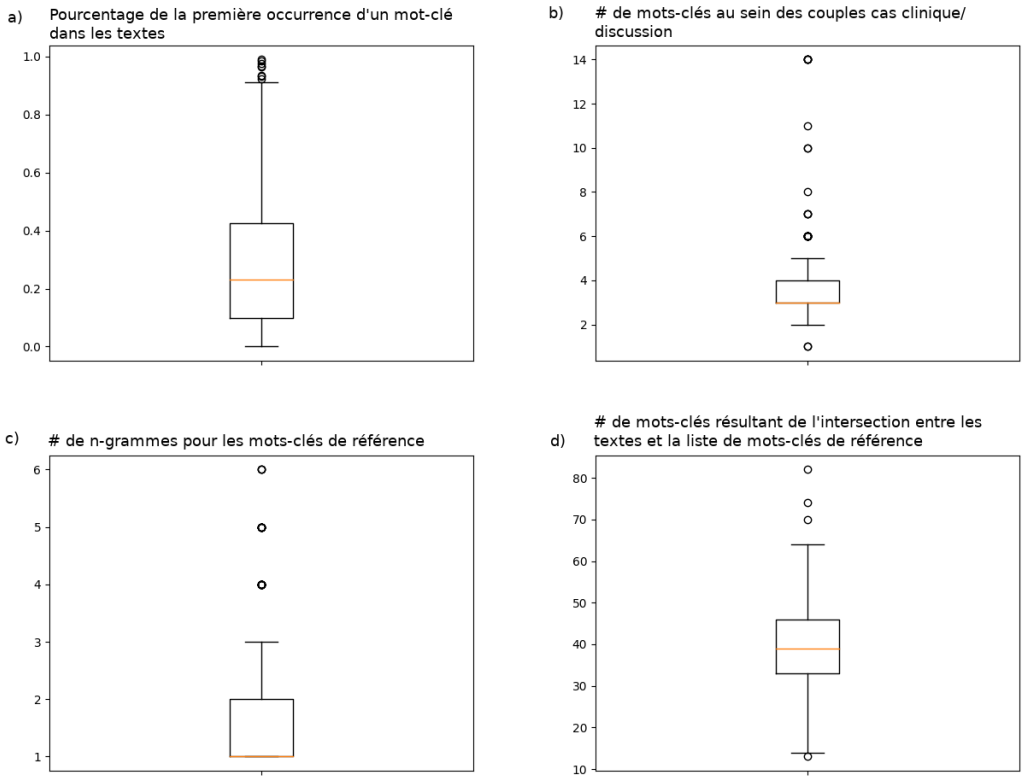

FIGURE 1 – Statistiques, présentées sous forme de *boxplot*, du corpus de textes et de mots-clés.

Ces dernières sont venues nourrir nos réflexions sur les méthodologies à privilégier pour la tâche 1. Nous discutons les différentes informations présentées dans cette figure :

a) la proportion de première occurrence d'un mot-clé au sein d'un cas clinique et d'une discussion. La plupart des premiers mots-clés qui apparaissent dans les textes apparaissent au début du texte.

b) le nombre de mots-clés à attribuer à chaque couple cas clinique/discussion. La plupart des couples sont annotés par 3 ou 4 mots-clés.

c) le nombre de grammes (mots) par mot-clé. Les couples sont essentiellement annotés par des unigrammes ou bi-grammes.

d) le nombre de mots-clés potentiels lorsque l'on considère l'intersection entre le vocabulaire contrôlé et les couples de textes. En médiane, nous observons qu'un couple contient des occurrences de 40 mots-clefs.

Nous avons aussi analysé le taux de recouvrement entre les mots-clés qui apparaissent dans les couples et les mots-clés attendus pour ces couples. Ce taux est calculé à partir de l'équation 1 dans laquelle $\mathcal{C}$ représente l'ensemble des couples cas clinique/discussion, M l'ensemble des mots-clés issus du vocabulaire contrôlé, T_c l'ensemble des mots du couple cas clinique/discussion c et K_c l'ensemble des mots-clés associés au couple cas clinique/discussion c.

$$R = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{|(M \cap T_c) \cap K_c|}{|K_c|} \quad (1)$$

Le taux de recouvrement global R des mots-clés attendus est de 0,72 *i.e.* 72% des mots-clefs attendus dans le jeu d'entraînement apparaissent dans les couples correspondants. Ainsi, dans le meilleur des cas, la MAP (cf. sous-section 1.1.2) pouvant être obtenue à l'aide d'une approche par extraction sera de 0,72 pour le jeu d'entraînement - un score pouvant être considéré comme respectable lorsque la tâche d'annotation est complexe. En tenant compte de l'observation d) proposée ci-dessus - en médiane, un couple contient des occurrences de 40 mots-clefs -, nous notons qu'il est possible de réduire la complexité de la tâche d'annotation en la redéfinissant comme une tâche d'extraction (au prix d'une réduction des performances théoriques maximales pouvant être atteintes, acceptable si la tâche d'annotation s'avère complexe). En la redéfinissant comme telle, il s'agit de définir une approche qui distinguera les mots-clefs pertinents parmi ceux observés dans les couples. Cette approche par extraction a l'avantage de réduire l'espace de recherche de manière significative : lors de l'annotation d'un couple, l'espace de recherche composé initialement de 1311 mots-clés est réduit à 40 mots-clés en médiane par couple. Les nombreux tests préliminaires effectués sur cette tâche sans réduction de l'espace de recherche, *i.e.* sans se restreindre à une approche par extraction, avec des stratégies multiples (statistique et par apprentissage), nous ont permis d'apprécier la potentielle difficulté de la tâche, et nous ont alors amené à concentrer nos efforts sur des approches par extraction. Ceux-ci seront détaillés par la suite.

1.1.2 Métrique d'évaluation

La métrique d'évaluation de cette première tâche est la MAP (*Mean Average Precision*). C'est une métrique populaire dans le domaine de la recherche d'information (Yue *et al.*, 2007). Chaque entrée est représentée par un vecteur binaire p, qui est l'espace des mots-clés de référence où la pertinence d'un mot-clé est symbolisée par un 1, et par un vecteur $\hat{p}$ dans lequel les mot-clés sont classés du plus pertinents au moins pertinent. Par conséquent, le classement des mots-clés récupérés a une incidence sur le score final. L'équation 2 formalise la modalité de calcul de la MAP.

$$MAP(p, \hat{p}) = \frac{1}{rel} \sum_{j:p_j=1} Prec@j \tag{2}$$

où $rel = |i : p_i = 1|$ est le nombre de mots-clés attendus et $Prec@j$ est le nombre de mots-clés pertinents dans les j premiers mots-clés issus du vecteur $\hat{p}$. À noter que lors des années précédentes du défi DEFT (2012 et 2016) la F1-mesure avait été utilisée.

1.2 Méthodologies et résultats

La méthodologie proposée repose sur 3 principales phases : i) un pré-traitement sur l'ensemble des mots-clés M issus du vocabulaire contrôlé et sur les textes des couples cas clinique/discussion, ii) le calcul d'un score de pondération TF-IDF des mots-clés de M pour chaque couple cas clinique/discussion et iii) le classement des mots-clés suite à une étape de post-traitement tenant compte des coefficients calculés lors de la précédente phase. Ces 3 étapes sont détaillées par la suite.

1.2.1 Pré-traitement des données

De nombreux termes au sein de la liste des mots-clés sont lexicalement proches et sémantiquement identiques, tels que « urètre » et « urèthre ». Ces termes ne sont cependant pas considérés comme similaires lors de la phase d'évaluation dû à leur différence orthographique, bien qu'ils représentent de manière implicite une même entité conceptuelle. En vue de réduire tant que possible la variabilité des observations lors de la phase de calcul des coefficients de pondération TF-IDF, et cela sans induire une perte dommageable d'information, nous avons souhaité adopter une approche permettant de considérer de tels termes comme identiques.

Un premier traitement est appliqué à l'ensemble des unigrammes extraits des éléments de M (mots-clés du vocabulaire contrôlé). La ponctuation, les chiffres ainsi que les *stopwords* ont été supprimés. Chaque unigramme est lemmatisé et racinisé pour obtenir un ensemble U d'unigrammes normalisés.

Dans l'objectif de regrouper les unigrammes proches sémantiquement (*e.g* « urètre » et « urèthre »), un second traitement est appliqué. Un nouvel ensemble U' d'unigrammes est construit à l'aide d'une mesure de similarité sémantique appliquée sur les éléments de U. Pour chaque couple d'unigrammes $(u_i, u_j) \in U^2$ avec $i \neq j$, les radicaux r_{u_i} et r_{u_j} sont extraits en soustrayant un préfixe et un suffixe communs à u_i et u_j, s'ils existent dans des listes prédifinies [1]. Une distance cosinus $cos(Emb(r_{u_i}), Emb(r_{u_j}))$ est ensuite calculée, avec $Emb(.)$ la fonction vecteur de plongement sémantique. Si $cos(Emb(r_{u_i}), Emb(r_{u_j})) > \lambda_0$ avec $\lambda_0 \in \mathbb{R}^+$, u_i et u_j sont considérés comme similaires, et dans ce cas seul l'unigramme le plus court est conservé. L'intérêt des préfixes et des suffixes est de limiter la similarité sémantique d'unigrammes spécifiques au domaine étudié. En effet, si deux unigrammes (u_i, u_j) ont pour suffixe « sarcome », tels que « liposarcome » et « carcinosarcome », $cos(Emb(u_i), Emb(u_j))$ sera proche de 1 bien qu'une distinction entre les deux soit nécessaire, alors que $cos(Emb(\text{"}lipo\text{"}), Emb(\text{"}carcino\text{"}))$ sera plus faible. Les vecteurs caractérisants le plongement sémantique de chaque unigramme, obtenus avec la méthode de Bojanowski *et al.* (2017) [2], correspondent à une moyenne pondérée de n-grammes de caractères, et sont donc particulièrement adaptés au traitement de racines d'unigrammes. Enfin, une abstraction M' de la liste des mots-clés de référence M est construite en substituant les unigrammes des éléments de M par ceux de U'.

L'ensemble des mots composants les cas cliniques et les discussions bénéficient d'un traitement similaire à celui appliqué aux mots-clés. Chaque couple est représenté par la concaténation du texte normalisé du cas clinique et de la discussion. Seuls les unigrammes présents dans U' sont conservés. Certains unigrammes ne sont caractéristiques que d'un seul mot-clé, tels que « *escherichia* » qui ne peut former que le mot-clé « *escherichia coli* ». Afin de pouvoir attribuer un coefficient de pondération TF-IDF non nul à ces mots-clés lorsqu'ils ne sont que partiellement observés, nous substituons ces unigrammes par la succession des unigrammes formant le seul mot-clé qu'ils peuvent constituer. La même procédure est appliquée aux bi-grammes ne pouvant former qu'un seul n-gramme strictement supérieur à 2.

1. Liste des préfixes utilisés : *acetyl, acetal, ana, anti, angio, antibiot, ante, ben, meth, eth, prop, but, pent, hex, hept, di, tri, tetra, carboxy, sulf, alca, hyper, hypo, cardio, psych, poly, pneumo, myco, meso, lymph, intra, hydro, immun, homo, endo, dys, chondro, met, micro, osteo, retro, hemangio*. Liste des suffixes utilisés : *tom, plast, scop, graph, oid, sarcom, log, om*.

2. Des vecteurs pré-entraînés ont été utilisés, disponibles sur `https://github.com/facebookresearch/fastText/blob/master/docs/pretrained-vectors.md`

1.2.2 Estimation du coefficient de pondération pour chaque mot-clé

L'objectif de cette phase est, pour chaque couple $c \in \mathcal{C}$, d'attribuer un coefficient de pondération pour chaque mot-clé de M'. A cette fin, un coefficient de pondération TF-IDF est calculé en tenant compte des n-grammes de rang 1 à 5 (cf. équation 3). Seuls les n-grammes présents dans M' sont conservés.

$$tfidf(t, c) = idf(t) \times (1 + log\ \ tf(t, c)) \tag{3}$$

où $tfidf(t, c)$ est le coefficient de pondération TF-IDF du n-gramme t pour le couple c, $tf(t, c)$ la fréquence du n-gramme t au sein du couple c et $idf(t)$ la fréquence inverse de document de t calculé selon l'équation 4.

$$idf(t) = 1 + log\frac{1 + |\mathcal{C}|}{1 + df(t)} \tag{4}$$

où $df(t)$ représente le nombre de couples $c \in \mathcal{C}$ contenant le n-gramme t.

Enfin, étant donné que certains n-grammes sont introduits par l'intermédiaire d'une substitution d'unigrammes (*e.g* « *escherichia coli* » substitue « *escherichia* »), leur fréquence est pondérée afin de marquer leur partielle observation (cf. équation 5).

$$tf(t, c) = entier(tf(t, c) \times \lambda_1) \tag{5}$$

où $entier(.)$ représente la fonction partie entière et $\lambda_1 \in [0, 1]$.

1.2.3 Post-traitement et classement des mots-clés

Dans l'objectif d'améliorer la MAP, un post-traitement est appliqué sur les coefficients de pondération TF-IDF obtenus précédemment. Dans un premier temps, les mots-clés de K, avec K mots-clés indexant les couples cas clinique/discussion dans le jeu d'entraînement, sont favorisés. Cela se traduit par l'équation 6.

$$tfidf(t, c) = tfidf(t, c) \times (1 + freq(t) \times \lambda_2) \tag{6}$$

où $freq(t)$ est la fréquence d'occurrence du terme t dans K et $\lambda_2 \in \mathbb{R}^+$.

Dans un second temps en observant le jeu d'entraînement, la tendance semble être que les n-grammes porteurs de l'information la plus spécifique sont privilégiés. Par exemple, si « tumeur » et « tumeur du rein » semblent pertinents pour indexer un couple, « tumeur du rein » est généralement favorisé. A cette fin, 3 stratégies sont appliquées de manière séquentielle sur les mots-clés à classer :

— $\forall(t_i, t_j) \in M'^2$ avec $i \neq j$, le coefficient de pondération TF-IDF w_i de t_i est incrémenté de $w_j \times \lambda_3$ pour chaque unigramme dans $t_i \cap t_j$, avec $\lambda_3 \in \mathbb{R}^+$.

— $\forall(t_i, t_j) \in M'^2$ avec $i \neq j$, si t_i est un n-gramme de rang supérieur à t_j, et $t_i \cap t_j \neq \{\varnothing\}$ et $w_i - w_j < max(w_i, w_j) \times \lambda_4$, avec $\lambda_4 \in \mathbb{R}^+$, alors $w_i := max(w_i, w_j)$ et w_j n'est plus candidat à l'indexation du couple cas clinique/discussion à traiter.

— $\forall(t_i, t_j) \in M'^2$ avec $i \neq j$, si $t_i \cap t_j \neq \{\varnothing\}$ et $w_i - w_j > max(w_i, w_j) \times \lambda_5$, avec $\lambda_5 \in \mathbb{R}^+$, alors w_j n'est plus candidat à l'indexation du couple cas clinique/discussion à traiter.

Enfin, pour chaque couple cas clinique/discussion c, la version non abstraite dans M des k_c mots-clés abstraits ayant le meilleur coefficient de pondération suite aux précédents traitements est utilisée comme indexe, où k_c représente le nombre de mots-clés attendus pour l'indexation du couple c. Cependant le processus de transformation d'un élément de M en un élément de M' n'est pas une application injective. Par exemple, « cancer de la prostate » $\in M$ et « cancer de prostate » $\in M$ correspondent au même élément normalisé « *canc prost* » $\in M'$. La stratégie de correspondance vers les versions non abstraites des mots-clés revient à utiliser l'élément de M correspondant à l'élément de M' sélectionné le plus représenté dans K. Lorsqu'il est impossible de départager les candidats, un choix par ordre alphabétique est effectué.

La MAP obtenue sur le jeu d'entraînement est de 0,42 avec $\lambda_0 = 0,6$, $\lambda_1 = 0,33$, $\lambda_2 = 110$, $\lambda_3 = 0,15$, $\lambda_4 = 0,25$ et $\lambda_5 = 0,45$. La MAP obtenue sur le jeu d'évaluation est de 0,40 avec ces mêmes paramètres. Sans limitation sur le nombre de mots-clés k_c à renvoyer pour l'indexation de chaque document c, et en ne considérant que les mots-clés dont le coefficient de pondération est non nul, la MAP obtenue est de 0,48.

2 Tâche 2 : Similarité sémantique entre les cas cliniques et les discussions

La tâche d'appariement textuel est la seconde tâche du défi DEFT 2019. Une tâche similaire avait été proposée lors du défi DEFT 2011 sur un corpus constitué de revues en Sciences Humaines et Sociales (Grouin *et al.*, 2011). Initialement, les organisateurs imaginaient une tâche associée au résumé automatique de textes mais de part les questions sous-jacentes liées à la complexité à évaluer une telle tâche (qu'est ce qui constitue un résumé de référence ? Comment évaluer la pertinence d'un résumé ?), ils l'ont transformée en une tâche d'appariement entre résumé et contenu d'article scientifique. Ils partent de l'hypothèse qu'un module de résumé textuel doit être en capacité d'évaluer le degré d'association entre le contenu d'un article et son résumé. Lors de cette précédente édition, deux phases de test avaient été réalisées. Ces phases se différenciaient sur le contenu des articles ; la première conservait la globalité de l'article et la seconde en supprimait l'introduction et la conclusion. Cette suppression part du principe qu'un même auteur aura tendance a paraphraser le résumé au travers de ces deux sections. Les résultats obtenus lors de cette tâche avaient été particulièrement bon : quatre équipes ont obtenu le score maximal lors de la première phase et deux d'entre elles réitèrent ce même score lors de la seconde phase.

2.1 Contexte de la tâche

2.1.1 Métrique d'évaluation

La métrique d'évaluation est la même que la précédente édition *i.e.* une évaluation binaire des résultats dans laquelle chaque prédiction exacte équivaut à 1 sinon à 0. Mis en formule, pour chacun des N cas cliniques r_i, le score $s(a_p(r_i), a_r(r_i))$ donné à chaque prédiction vaut 0 ou 1 selon que la discussion prédite $a_p(r_i)$ est, ou pas, la discussion de référence $a_r(r_i)$ (cf. équation 7).

$$s(a_p(r_i), a_r(r_i)) = \begin{cases} 1 & \text{si } a_p(r_i) = a_r(r_i) \\ 0 & \text{sinon.} \end{cases} \quad (7)$$

Le score global correspond à la moyenne des scores obtenus sur l'ensemble des prédictions (cf. équation 8).

$$S(p) = \frac{1}{N} \sum_{i=1}^{N} s(a_p(r_i), a_r(r_i)) = \frac{1}{N} \sum_{i=1}^{N} |r_i; a_p(r_i) = a_r(r_i)| \quad (8)$$

2.1.2 Méthodes proposées lors de DEFT 2011

Plusieurs méthodes proposées lors de DEFT 2011 ont obtenu le score maximal au moins pour la phase 1. C'est le cas par exemple de Hoareau *et al.* (2011) ayant obtenu 100% et 99,5% respectivement à la phase 1 et 2. Les auteurs proposent de modéliser chaque document au sein d'espaces sémantiques construits au travers de projections aléatoires par le biais d'une méthodologie intitulée *Random Indexing*. Ils calculent ensuite une matrice de distances en exploitant la distance euclidienne pondérée entre chaque document de l'espace sémantique. Cette matrice, modélisant un graphe à N noeuds et N^2 arcs, permet aux auteurs de construire un graphe biparti dans lequel un article est associé au résumé le plus proche. En cas d'ambiguïté, *i.e.* plusieurs articles pour un même résumé, une procédure itérative basée sur la minimisation des distances entre les différents articles concernés est mise en place. Un autre exemple d'utilisation des espaces sémantiques a été proposé dans le cadre d'une approche supervisée (Bestgen, 2011). Cette approche a obtenu 100% lors des deux phases de test. Les espaces sémantiques sont obtenus ici au travers d'une approche de réduction matricielle basée sur une SVD (*Singular Value Decomposition*) : la méthode LSA (*Latent Semantic Analysis*). L'algorithme d'apprentissage utilisé est un SVM (*Support Vector Machine*) qui tient compte des espaces sémantiques comme caractéristiques pour l'apprentissage. Le SVM exploite une approche multi-classes ainsi qu'une stratégie d'appariement par le meilleur d'abord en tenant compte des valeurs de décision de la procédure SVM comme un score de comptabilité avec chacune des catégories et donc avec chaque article. Enfin, un dernier exemple obtenant 100% et 90,9% sur la phase 1 et 2 propose une approche orientée autour des constructions lexicales (Lejeune *et al.*, 2011). Les auteurs emploient la méthode *rstr-max* permettant de détecter les chaînes de caractères répétées maximales entre les articles et les résumés. Cette méthode permet de définir une notion d'affinité caractérisée par la taille en caractères de la plus grande chaîne de caractères et le nombre total de chaînes de caractères en commun pour chaque couple potentiel. Par le biais de ces affinités, le résumé adéquat pour un article sera celui qui partagera le plus grand nombre d'affinités avec un article.

2.1.3 Spécificités du corpus de DEFT 2019

Les expérimentations que nous avons menées ont permis de détecter des propriétés potentiellement indésirables du jeu de données. En effet, ce dernier référence uniquement une seule discussion pour chacun des cas cliniques. Cependant, plusieurs discussions possèdent des copies identiques avec des identifiants différents. Par conséquent, le fait qu'un cas clinique ne référence pas toutes les copies de la discussion qui lui est rattachée génère des erreurs arbitraires lors de l'évaluation. Dans le cas présent, la mise en place d'une stratégie d'appariement (minimisation des distances, stratégie par le

meilleur d'abord, maximisation des affinités) ne permet pas de différencier l'identifiant exact parmi les différentes copies d'une même discussion par rapport à un cas clinique donné. Par conséquent, les méthodes décrites en sous-section 2.2 ne tiennent pas compte d'une stratégie d'appariement mais uniquement d'un tirage avec remise parmi l'ensemble des résumés possibles pour chaque cas clinique.

2.2 Méthodologies et résultats

Chaque approche expérimentée s'appuie sur des cas cliniques et des discussions lemmatisés par l'intermédiaire de l'outil *TreeTagger*. L'approche ayant servi de *baseline* repose sur une analyse des similarités lexicales partagées entre les textes par l'intermédiaire d'une adaptation de la similarité de Lin (Lin, 1998). Comme montré dans l'équation 9, la similarité entre un cas clinique s_1 et une discussion s_2 est calculée comme le logarithme de la probabilité de la somme pondérée des unigrammes en communs entre les deux textes, divisée par le logarithme de la probabilité de la somme pondérée de tous les mots dans les deux textes - formulation du coefficient de DICE basée sur des métriques proposées par la théorie de l'information).

$$sim(s_1, s_2) = \frac{2 \times \sum_{w \in s_1 \cap s_2} \log P(w)}{\sum_{w \in s_1} \log P(w) + \sum_{w \in s_2} \log P(w)} \tag{9}$$

Les probabilités $P(w)$ des unigrammes w sont estimées sur les données d'entraînement (Agirre *et al.*, 2016). Cette méthodologie obtient un score de 0,638.

Les deux prochaines approches ont été soumises lors de la phase de test. Elles diffèrent principalement sur la manière de représenter les cas cliniques et les discussions. La première approche exploite une méthode des *k* plus proches voisins avec une distance euclidienne (cf. tableau 2). Les cas cliniques et les discussions sont modélisés au travers d'une représentation vectorielle des valeurs de TF-IDF associées à leurs n-grammes [3].

Distances	Score
Euclidienne	**0,748**
Manhattan	0,141
Chebyshev	0,352
Hamming	0,010
Canberra	0,045
Braycurtis	0,741

TABLE 2 – Résultats issus de la méthode des *k* plus proches voisins en tenant compte de différentes distances.

Tandis que la seconde approche mise en place se base sur les espaces sémantiques des différents documents calculés à partir de la méthode LSA (cf. tableau 3) et d'une représentation vectorielle des textes qui tient compte des valeurs de TF-IDF associées aux n-grammes [4]. Ces espaces permettent ensuite de calculer une matrice de distances l^2 normalisée entre les cas cliniques et les discussions à

3. Les n-grammes considérés vont de l'unigramme au pentagramme.

4. La représentation vectorielle des textes au travers d'une valeur binaire de présence ou d'absence d'un terme a également été testée mais elle obtient un score moins performant.

partir d'une distance euclidienne. L'appariement est ensuite réalisée en minimisant la distance entre un cas clinique et les discussions.

Dimensions	Score
200	0,707
300	0,745
400	0,728
500	0,734
1000	**0,755**

TABLE 3 – Résultats en fonction du nombre de dimensions sélectionnées par l'intermédiaire de la méthode LSA.

Lors de la phase de test, l'approche obtenant le meilleur résultat est la méthode des *k* plus proches voisins utilisant la mesure euclidienne avec un score de 0,86. Toutefois, les organisateurs de la tâche ont recalculé les scores en tenant compte des doublons du corpus de test. L'actualisation du corpus a permis d'améliorer les performances de cette même approche avec un score de 0,91.

3 Conclusion

Ces travaux présente la contribution du LGI2P pour la tâche 1 et 2 du défi DEFT 2019. Ces tâches ont porté respectivement sur l'indexation et la similarité entre documents. La particularité de ce défi portait sur la nature biomédicale et clinique du jeu de données qui était constitué d'un ensemble de couples cas clinique/discussion. Les méthodologies développées pour ces deux tâches s'appuient sur des techniques provenant du domaine de la recherche d'information notamment au travers de l'utilisation du coefficient de pondération TF-IDF. Lors des phrases d'évaluation la meilleure approche pour la tâche 1 a obtenu un score 0,48, tandis que pour la tâche 2 la meilleure approche a obtenu un score de 0,91.

Références

AGIRRE E., BANEA C., CER D., DIAB M., GONZALEZ-AGIRRE A., MIHALCEA R., RIGAU G. & WIEBE J. (2016). Semeval-2016 task 1 : Semantic textual similarity, monolingual and cross-lingual evaluation. p. 497–511 : International Workshop on Semantic Evaluation.

BESTGEN Y. (2011). Lsvma : au plus deux composants pour apparier des résumés à des articles. p. 105–114 : Actes du septième DÉfi Fouille de Textes.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching word vectors with subword information. volume 5, p. 135–146.

CHARTIER J.-F., FOREST D. & LACOMBE O. (2016). Alignement de deux espaces sémantiques à des fins d'indexation automatique. p. 13–19 : Actes du douxième DÉfi Fouille de Textes.

DEERWESTER S., DUMAIS S. T., FURNAS G. W., LANDAUER T. K. & HARSHMAN R. (1990). Indexing by latent semantic analysis. In *Journal of the American Society for Information Science*, p. 391–407.

FIORINI N., RANWEZ S., MONTMAIN J. & RANWEZ V. (2015). Usi : a fast and accurate approach for conceptual document annotation. In *BMC Bioinformatics*, p. 1471–2105.

GRABAR N., GROUIN C., HAMON T. & CLAVEAU V. (2019). : Actes du quinzième DÉfi Fouille de Textes.

GROUIN C., FOREST D., PAROUBEK P. & ZWEIGENBAUM P. (2011). Présentation et résultats du défi fouille de texte deft2011. p. 3–14 : Actes du septième DÉfi Fouille de Textes.

GUPTA V. & LEHAL G. S. (2010). A survey of text summarization extractive techniques. In *Journal of emerging technologies in web intelligence*, p. 258–268.

HAMDAN H. (2015). Sentiment analysis in social media. In *P.h.D thesis at Aix-Marseille*, p. 165.

HARISPE S., SÁNCHEZ D., RANWEZ S., JANAQI S. & MONTMAIN J. (2014). A framework for unifying ontology-based semantic similarity measures : A study in the biomedical domain. In *Journal of biomedical informatics*, p. 38–53.

HOAREAU Y. V., AHAT M., PETERMANN C. & BUI M. (2011). Couplage d'espaces sémantiques et de graphes pour le deft 2011 : une approche automatique non supervisée. p. 115 : Actes du septième DÉfi Fouille de Textes.

JONES K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. In *Journal of documentation*, p. 11–21.

KIM T.-Y., KIM J., LEE J. & LEE J.-H. (2014). A tweet summarization method based on a keyword graph. In *Conference on Ubiquitous Information Management and Communication*, p.9̃6 : ACM.

LEJEUNE G., BRIXTEL R. & GIGUET E. (2011). Deft 2011 : appariements de résumés et d'articles scientifiques fondés sur des distributions de chaînes de caractères. p. 53–64 : Actes du septième DÉfi Fouille de Textes.

LIN D. (1998). An information-theoretic definition of similarity. volume 98, p. 296–304 : Proceedings of the Fifteenth International Conference on Machine Learning.

LITVAK M. & LAST M. (2008). Graph-based keyword extraction for single-document summarization. In *Multi-source Multilingual Information Extraction and Summarization*, p. 17–24 : Association for Computational Linguistics.

MAHATA D., KURIAKOSE J., SHAH R. R. & ZIMMERMANN R. (2018). Key2vec : Automatic ranked keyphrase extraction from scientific articles using phrase embeddings. In *Association for Computational Linguistics*.

MARCHAND M., FOUQUIER G., MARCHAND E. & PITEL G. (2016). Représentation vectorielle de documents pour l'indexation de notices bibliographiques. p. 34–40 : Actes du douxième DÉfi Fouille de Textes.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, p. 3111–3119.

NALLAPATI R., ZHOU B., GULCEHRE C., DOS SANTOS C. & XIANG B. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Conference on Computational Natural Language Learning*.

NANAS N., UREN V. & ROECK A. D. (2003). A comparative study of term weighting methods for information filtering. In *Knowledge Media Institue*.

WANG R., LIU W. & MCDONALD C. (2014). Corpus-independent generic keyphrase extraction using word embedding vectors. In *Software Engineering Research Conference*, volume 39.

WITTEN I. H., PAYNTER G. W., FRANK E., GUTWIN C. & NEVILL-MANNING C. G. (2005). Kea : Practical automated keyphrase extraction. In *Design and Usability of Digital Libraries : Case Studies in the Asia Pacific*, p. 129–152 : IGI Global.

YUE Y., FINLEY T., RADLINSKI F. & JOACHIMS T. (2007). A support vector method for optimizing average precision. In *SIGIR*, p. 271–278 : ACM.

ZHANG C. (2008). Automatic keyword extraction from documents using conditional random fields. In *Journal of Computational Information Systems*, p. 1169–1180.