Les espaces sémantiques de mots-clés : une méthode d'indexation automatique de documents par assignation de mots-clés

Keyword Representations in Semantic Vector Space: a Keyword Assignment Method for Automatic Document Indexing

Jean-François Chartier[1], Dominic Forest[1]

[1] École de Bibliothéconomie et des Sciences de l'Information, Université de Montréal, Canada.

RÉSUMÉ. Avec la croissance extrêmement rapide de la quantité de documents numériques dans nos sociétés, l'automatisation de l'indexation est devenue un enjeu de recherche central pour la gestion documentaire. Plusieurs compétitions scientifiques portant sur des tâches d'indexation automatique ont vu le jour ces dernières années. Cet article rend compte de notre participation à l'une d'entre elles, soit l'édition 2016 du Défi fouille de textes (DEFT-2016). Dans un premier temps, nous présentons un état de la situation concernant l'importance, mais aussi les enjeux et les défis de l'indexation automatique. Après avoir présenté les grandes lignes de la campagne d'évaluation DEFT-2016, nous introduisons l'approche que nous avons développée. Celle-ci repose sur la construction d'un espace sémantique de mots-clés. L'évaluation des performances de notre approche et l'analyse des résultats suggèrent que notre méthode est particulièrement adaptée à des tâches d'indexation automatique qui nécessitent une part importante d'assignation de mots-clés contrôlés qui sont absents du contenu textuel des documents.

ABSTRACT. With the extremely rapid growth of the amount of digital documents in our societies, automatic keyword indexing has become a central research issue in information retrieval and document management. Several scientific competitions dealing with automatic indexing tasks have emerged in recent years. This article reports our participation in one of them, the 2016 edition of Défi Fouille de Texte (DEFT-2016). First, we present a state of the art regarding the importance, the issues and the challenges of automatic keyword indexing. After presenting the context and the task of the DEFT-2016, we introduce the method we have developed. This method is based on the construction of a keyword semantic vector space. The evaluation of our method and the analysis of the results suggest that our approach is particularly adapted to automatic keyword indexing tasks which require a large proportion of controlled keyword assignment that are absent from the text content of the documents.

MOTS-CLÉS. Indexation automatique, Assignation de mots-clés, Extraction de mots-clés, Algorithme non-supervisé, Algorithme supervisé, Espace sémantique, Défi fouille de textes, DEFT.

KEYWORDS. Automatic Keyword Indexing, Keyword Assignment, Keyword Extraction, Supervised Machine Learning, Unsupervised Machine Learning, Semantic Vector Space, Défi Fouille de Textes, DEFT.

Introduction

L'indexation est une activité centrale dans la gestion documentaire. Elle consiste à décrire le contenu d'un document, ses thèmes ou ses sujets, à l'aide de mots-clés et ce en vue d'une recherche ultérieure de l'information contenue dans ce document. Ce travail est traditionnellement réalisé manuellement par des documentalistes professionnels, mais avec la croissance constante de la quantité de documents diffusés en format numérique, cette pratique est devenue impossible à poursuivre. À titre d'exemples, la base de données bibliographiques Web of Science contient aujourd'hui plus de 90 millions de notices [THEC17], la base de données Pascal et Francis contient plus de 14 millions de notices [BASE00] et MEDLINE contient plus de 23 millions de notices. La quantité de documents est de plus en plus importante. En 1996, MEDLINE ajoutait environ 300 000 nouvelles notices par année à sa base de données. Aujourd'hui, c'est environ 900 000 nouvelles notices qui s'ajoutent chaque année [MEDL16]. Sachant que le coût estimé pour l'indexation d'un document par des documentalistes professionnels chez MEDLINE est d'environ 9,40$ [LPWZ15], l'accroissement rapide de la quantité de documents numériques est associé à un fardeau économique immense. Dans ce contexte, non seulement l'automatisation de l'indexation est une problématique scientifique de premier

plan, mais elle est également devenue un enjeu économique majeur dans l'industrie de la gestion documentaire.

Dans la communauté scientifique, la problématique de l'automatisation de l'indexation des documents mobilise les chercheurs depuis plus de trente ans [JONE72, SAWY75, VANR79]. Toutefois, c'est surtout depuis une dizaine d'années que nous observons une forte intensification de la recherche dans ce domaine. C'est ce qu'indique du moins une analyse de la fréquence des publications sur cette problématique. La quantité d'articles scientifiques touchant cette problématique a augmenté très rapidement. La figure 1 illustre l'évolution cumulative de ces publications dans la base de données bibliographique SCOPUS. En 2016, ces publications totalisaient 1 007 références.[1]

Le même phénomène s'observe dans l'industrie comme en témoigne l'évolution du nombre de brevets délivrés sur des technologies d'indexation automatique. La figure 1 illustre l'évolution cumulative de ces brevets dans la base de données The Lens. En 2016, ce nombre totalisait 1 029 brevets, avec une courbe de croissance très similaire à celle des articles scientifiques.[2]

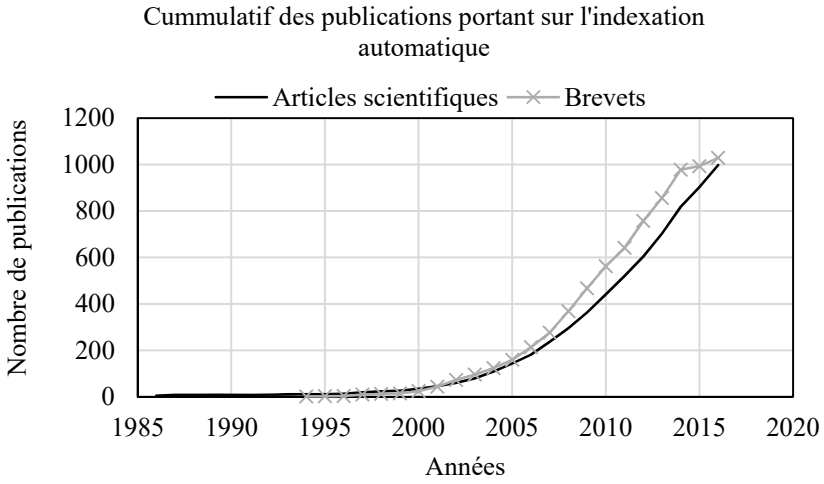

Figure 1. *Évolution cumulative des publications portant sur l'indexation automatique.*

L'apparition de plusieurs compétitions scientifiques (ou campagnes d'évaluation) portant sur les enjeux de l'indexation automatique s'avère un autre indicateur de l'importance de cette problématique. Parmi ces compétitions, on retrouve la campagne SemEval de 2010 [KMKB10], les campagnes 2012 et 2016 de DEFT [DAIL16, PZFG12] et les campagnes annuelles de BIOASQ [TBMP15] organisées par PubMed/MEDLINE. Ces compétitions sont des moments importants permettant d'établir l'état de l'art dans un domaine qui évolue très rapidement. Elles s'avèrent en effet des cadres rigoureux et transparents grâce auxquels peuvent être comparées et évaluées - sur la base d'un même ensemble de paramètres - plusieurs méthodes alternatives.

Le but de cet article est de présenter notre participation à l'édition 2016 de la compétition Défi Fouille de Textes (DEFT 2016). Un premier article court a été publié en 2016 dans les Actes du colloque JEP-TALN 2016 où la méthode que nous avions développée et les résultats obtenus lors de notre participation à la compétition étaient sommairement présentés [CHFL16]. Dans cette version longue, nous visons 3 objectifs. Premièrement, nous présentons un état de l'art de l'indexation automatique, basé sur une typologie des mots-clés et une typologie des méthodes d'indexation automatique (section 1). Dans un deuxième temps, nous présentons en détail les modalités de notre participation à la campagne d'évaluation DEFT 2016, ainsi que la méthode que nous avons développée

[1] Il s'agit d'une estimation conservatrice obtenue avec la requête : (TITLE-ABS-KEY("keyphrase extraction") or TITLE-ABS-KEY("keyword extraction") or TITLE-ABS-KEY("keyterm extraction") or TITLE-ABS-KEY("keyphrase assignment") or TITLE-ABS-KEY("keyword assignment") or TITLE-ABS-KEY("keyterm assignment")).

[2] Estimation effectuée sur lens.org avec la requête : (CLAIMS:("keyphrase extraction") || CLAIMS:("keyword extraction") || CLAIMS:("keyterm extraction") || CLAIMS:("keyphrase assignment") || CLAIMS:("keyterm assignment") || CLAIMS:("keyword assignment")).

dans ce contexte (section 2). Enfin, nous apportons des éléments d'explication aux forces et aux faiblesses de notre méthode (section 3).

1. État de l'art

1.1. *Typologie des mots-clés*

On distingue généralement deux types de mots-clés, soit les mots-clés contrôlés et les mots-clés non-contrôlés. Les premiers appartiennent à un langage documentaire prédéfini, tel un thésaurus. Ce sont généralement des mots-clés métiers ou des taxinomies spécifiques à des domaines de spécialité (domaine médical, scientifique, légal ou autre). Ce sont des mots-clés normalement assignés à une notice par les documentalistes professionnels ou des experts du domaine. Indexer des documents à l'aide du thésaurus MeSH (Medical Subject Headings) est un exemple d'indexation contrôlée, car elle est circonscrite aux entrées de ce langage documentaire.

Les mots-clés non-contrôlés n'appartiennent quant à eux à aucun langage documentaire prédéfini. Il n'y a, en principe, aucune contrainte dans le choix des mots-clés d'indexation d'un document. Dans la pratique, même les mots-clés non-contrôlés sont soumis à des contraintes informelles et des règles de bonnes pratiques. Ils doivent par exemple être pertinents, non-redondants et représentatifs du contenu du document indexé. Mais il y a aussi plusieurs autres types de contraintes externes au document indexé comme des contraintes sociales. C'est le cas par exemple de la folksonomie de Wikipedia. Les pages de Wikipedia font l'objet d'une indexation non-contrôlée qui est néanmoins soumise à des dynamiques sociales spécifiques qui déterminent le choix des mots-clés [SSGS12].

À cette première distinction (entre les mots-clés contrôlés et non-contrôlés) s'en ajoute une seconde, entre les mots-clés présents dans les documents indexés et les mots-clés absents des documents indexés. Les premiers sont des mots-clés que l'on retrouve dans le contenu textuel d'un document. Ce sont des mots-clés souvent présents soit dans le titre ou dans le résumé du document. L'occurrence de ces mots-clés peut être exacte ou partielle. Elle est exacte lorsque la graphie du mot-clé est présente telle quelle dans le document et elle est partielle lorsque l'une de ses flexions s'y retrouve, par exemple sa forme racinée ou lemmatisée.

Les mots-clés absents ne sont pas présents dans le contenu textuel du document indexé, on ne retrouve ni leur graphie exacte ni l'une de leurs flexions. Ce sont des mots-clés inférés à partir du contenu d'un document. Ces mots-clés inférés vont généralement entretenir différentes relations sémantiques avec le contenu textuel du document indexé. Ils sont, par exemple, des synonymes, des hyperonymes, des hyponymes ou des méronymes des autres mots présents dans le document indexé.

Selon cette double distinction, nous nous retrouvons donc avec quatre types de mots-clés d'indexation. Cette typologie est illustrée dans le tableau 1.

	Mots-clés contrôlés	**Mots-clés non-contrôlés**
Mots-clés présents	Mots-clés contrôlés présents dans le document.	Mots-clés non-contrôlés présents dans le document.
Mots-clés absents	Mots-clés contrôlés absents du document.	Mots-clés non-contrôlés absents du document.

Tableau 1. *Typologie des mots-clés.*

Les forces et les faiblesses des indexations à base de mots-clés contrôlés et non-contrôlés demeurent l'objet de plusieurs débats [FRQU07]. L'usage de mots-clés contrôlés fait partie de la pratique

traditionnelle des documentalistes professionnels. On retrouve généralement des indexations à base de mots-clés contrôlés dans des secteurs d'activités spécialisés où des professionnels, comme des bibliothécaires et des juristes, sont spécialement formés à l'utilisation d'un vocabulaire contrôlé pour la recherche d'information. On retrouve habituellement des indexations à base de mots-clés non-contrôlées dans des secteurs d'activités génériques où la recherche d'information se fait via un moteur de recherche plein texte (texte libre), par exemple sur le Web.

Lorsque l'indexation est basée sur des mots-clés contrôlés, il est courant que ceux-ci ne soient pas présents dans le contenu textuel du document indexé. Ceci constitue d'ailleurs l'une des principales motivations pour l'utilisation d'un vocabulaire contrôlé. Il en va autrement lorsque l'indexation est basée sur des mots-clés non-contrôlés, car plusieurs moteurs de recherche plein texte sont encore des moteurs de type booléen qui ne peuvent pas inférer de relations sémantiques entre le contenu d'un document et les entrées d'un thésaurus.

1.2. *Typologie des méthodes d'indexations automatiques*

La prédiction des mots-clés contrôlés et des mots-clés non-contrôlés peut être réalisée en déployant différentes méthodes d'indexation automatique. Ces méthodes peuvent être classées selon deux axes. Le premier axe distingue les méthodes d'extraction et les méthodes d'assignation de mots-clés. L'indexation par extraction consiste à indexer un document uniquement à l'aide de mots-clés extraits de son contenu textuel. Ce sont donc des méthodes qui peuvent uniquement indexer un document via des mots-clés présents dans le document. L'indexation par assignation attribue à un document des mots-clés issus d'un langage documentaire contrôlé, que ceux-ci soient présents ou non dans le document.

Le deuxième axe distingue les algorithmes supervisés et les algorithmes non-supervisés. Les algorithmes non-supervisés sont mobilisés par des méthodes d'ordonnancement ou de regroupement, alors que les algorithmes supervisés sont mobilisés par des méthodes de classification. La plupart des méthodes d'indexation automatique peuvent être situées sur ces deux axes (axe 1 : extraction vs assignation et axe 2 : algorithmes supervisés vs algorithmes non-supervisés), elles seront soit des méthodes d'extraction basées sur des algorithmes non-supervisés, des méthodes d'extraction basées sur des algorithmes supervisés, des méthodes d'assignation basées sur des algorithmes non-supervisé ou des méthodes d'assignation basée sur des algorithmes supervisés. Cette typologie est illustrée dans le tableau 2.

	Algorithme non-supervisé	**Algorithme supervisé**
Méthode d'extraction	Méthodes d'ordonnancement des mots présents dans un document.	Méthode de classification binaire des mots présents dans un document.
Méthode d'assignation	Méthode de tri des mots-clés d'un vocabulaire contrôlé présent dans un document.	Méthode de classification multi-étiquettes des documents.

Tableau 2. *Typologie des méthodes d'indexation automatique de mots-clés.*

1.2.1. *Extraction non-supervisée*

Les méthodes d'extraction qui mobilisent des algorithmes non-supervisés indexent un document avec des mots-clés non-contrôlés présents dans ce document. Le schéma général de ces méthodes est simple. Après avoir extrait d'un document un ensemble de mots-clés candidats, ceux-ci sont ensuite ordonnancés selon un ou plusieurs coefficients de pondération. Finalement, le sous-ensemble de mots-

clés candidats associés à une valeur de pondération qui dépasse un seuil donné est sélectionné pour l'indexation.

Dénotons par $d_i = \{t_{i1} \dots t_{in}\}$ un ensemble de n mots-clés candidats pour un document i et par $p(t_{ij})$ la pondération du candidat j pour le document i selon un coefficient donné. L'opération d'indexation d'un document par extraction non-supervisée correspond généralement à la fonction suivante :

$$\mathbb{I}_i^{EN} = \bigcup_{t_{ij} \in d_i} \{t_{ij} : p(t_{ij}) > min\} \quad [1]$$

En d'autres mots, tous les candidats associés à une pondération supérieure à un seuil donné sont sélectionnés pour former l'indexation du document. Une alternative courante consiste à sélectionner les k candidats d'un document qui ont les pondérations les plus élevées.

La pondération des mots-clés candidats est généralement basée sur des critères statistiques ou linguistiques. Les critères statistiques permettent de pondérer les mots-clés candidats selon leur degré de représentativité du contenu du document et leur degré de discrimination par rapport à d'autres documents. C'est par exemple ce que permettent de faire des coefficients distributionnalistes comme le TF-IDF et ses variantes [JONE72, SAWY75, SPAR74]. Une autre classe de pondération statistique regroupe des coefficients de connexités basés sur la cooccurrence, par exemple le coefficient PageRank [BEMM15, BOUG15, MITA04, WAXI08].

Des critères linguistiques permettent de pondérer les mots-clés candidats selon leur rôle linguistique dans le document. Une technique classique consiste à ordonnancer selon leur longueur les mots-clés candidats qui forment des syntagmes nominaux. Les syntagmes plus longs expriment généralement des concepts plus importants que les syntagmes courts et sont donc sélectionnés en premier [BACO00, ERCI07, KIBK10]. Les mots-clés candidats peuvent aussi être ordonnancés selon leur position dans le document. Les mots-clés candidats présents au début d'un document, soit dans le titre ou la première phrase du document, sont généralement plus importants que les candidats occurrents au milieu du document [NGKA07, WPFG99].

1.2.2. *Extraction supervisée*

Les méthodes d'extraction basées sur des algorithmes supervisés ont le même objectif que le type de méthodes précédent : elles cherchent à indexer un document avec des mots-clés non-contrôlés qui sont présents dans le document. Cependant, contrairement aux méthodes précédentes, elles ne sont pas basées sur une fonction d'ordonnancement, mais sur une fonction de classification binaire. Le schéma habituel consiste à extraire d'un document un ensemble de mots-clés candidats et ensuite à classer chaque candidat comme positif (candidat sélectionné) ou négatif (candidat non sélectionné).

Dénotons par $g: \mathbb{R}^m \longrightarrow \{1, -1\}$ une fonction de classification binaire qui prend en argument un vecteur de pondération $\mathbf{t}_{\mathrm{ij}} = \left(p(t_{ij})_1 \dots p(t_{ij})_m\right)$ et lui associe une valeur positive si le candidat $\mathbf{t}_{\mathrm{ij}}$ est sélectionné et une valeur négative si ce n'est pas le cas. L'opération d'indexation d'un document i par extraction supervisée peut être représentée formellement par la fonction suivante :

$$\mathbb{I}_i^{ES} = \bigcup_{t_{ij} \in d_i} \{t_{ij} : g(\mathbf{t}_{\mathrm{ij}}) = 1\} \quad [2]$$

Un vecteur de pondération est généralement composé de plusieurs coefficients pondérations statistiques et linguistiques. Par exemple, dans la méthode d'indexation KEA, les vecteurs de pondérations ont deux dimensions et sont composés du TF-IDF d'un candidat et sa position dans le document [WPFG99].

Dans les méthodes d'indexation par extraction supervisée, la fonction $g(\mathbf{t}_{ij})$ est approximée par apprentissage statistique sur une collection d'exemplaires $\{(\mathbf{t}_{ij}, y) \in \mathbb{R}^m \times \{1, -1\}\}$ construite généralement par des documentalistes professionnels. Cette base d'exemplaires sert de référence à partir de laquelle un algorithme est entraîné. L'apprentissage statistique consiste dans ce contexte à induire quels patrons de pondérations sont de bons prédicteurs qu'un candidat j constitue un mot-clé que des documentalistes professionnels sélectionnerait pour l'indexation d'un document i.

Plusieurs modèles d'apprentissage statistiques peuvent approximer $g(\mathbf{t}_{ij})$. Par exemple, la méthode KEA mobilise un algorithme d'apprentissage bayésien naïf [WPFG99]. L'algorithme des k plus proches voisins est aussi couramment mobilisé par ce type de méthode [WAXI08], des algorithmes génétiques et des arbres de décisions ont également été appliqués à ce type de tâches [TURN00] ainsi que des séparateurs à vastes marges [FPWG99].

1.2.3. *Assignation non-supervisée*

L'opérationnalisation la plus simple des méthodes d'assignation basées sur des algorithmes non-supervisés suit un schéma composé de deux étapes. Dans un premier temps, un ensemble de mots-clés est extrait d'un document via une méthode d'indexation par extraction non-supervisée. Dans un second temps, cet ensemble est comparé avec un vocabulaire contrôlé et seulement l'intersection de ces deux ensembles est retenue pour former l'indexation d'un document.

Si nous dénotons par $\mathbb{V}$ un vocabulaire contrôlé et par $t_{ij} \in \mathbb{I}_i^{EN}$ l'ensemble des mots-clés d'indexation extrait de manière non-supervisée, l'opération d'indexation par assignation non-supervisée peut être représentée de la manière suivante :

$$\mathbb{I}_i^{AN} = \bigcup_{t_{ij} \in \mathbb{I}_i^{EN}} \{t_{ij} : t_{ij} \subseteq \mathbb{V}\} \qquad [3]$$

La comparaison entre $\mathbb{I}_i^{EN}$ et $\mathbb{V}$ représente l'opération centrale de ce type de méthode. Elle se limite rarement à une comparaison stricte des chaînes de caractères. Différentes techniques de normalisation des graphies, principalement la racinisation, la lemmatisation et le filtrage avec un antidictionnaire, sont utilisées pour comparer différentes flexions d'un même mot ou syntagme.

Dans une variante de ce type de méthode, la comparaison de $\mathbb{V}$ avec $\mathbb{I}_i^{EN}$ est remplacée par une comparaison avec $\mathbb{I}_i^{ES}$, une indexation par extraction supervisée [JMDA12, MEWI06]. Toutefois, ceci ne change pas le caractère non-supervisé de l'opération d'assignation des mots-clés via un vocabulaire contrôlé. Cette variante remplace simplement une opération d'ordonnancement par une opération de classification.

Les méthodes d'assignation non-supervisées supposent que les mots-clés d'indexation assignés à un document sont présents dans le contenu textuel du document. C'est évidemment une limite importante de ce type de méthode, car les mots-clés contrôlés sont souvent absents des documents et doivent au contraire être inférés.

1.2.4. *Assignation supervisée*

Les méthodes d'assignation qui utilisent des algorithmes supervisés indexent un document avec des mots-clés appartenant à un vocabulaire contrôlé. Contrairement aux approches d'assignation non-supervisées, elles permettent d'indexer des documents avec des mots-clés contrôlés absents de son contenu. Ce sont des méthodes d'inférence statistique basées sur de la classification multi-étiquettes de documents [SORO10, TSKA06]. Autrement dit, dans ce type d'approche, chaque mot-clé d'un vocabulaire contrôlé représente une classe possible pour un document et l'opération d'indexation consiste à classer chaque document dans plusieurs de ces classes.

Même si elles sont toutes les deux basées sur des algorithmes de classification supervisés, l'extraction supervisée et l'assignation supervisée sont des opérations très différentes. Dans la première, tel que discuté précédemment, l'opération de classification correspond à la fonction $g: \mathbb{R}^m \longrightarrow \{1, -1\}$. Elle est une classification binaire, disjointe, d'un mot-clé candidat. Dans la seconde, l'opération de classification correspond à la fonction $\mathcal{G}: \mathbb{R}^n \longrightarrow \{1, -1\}^k$, qui prend en argument un vecteur $\mathbf{d}$ modélisant le contenu textuel d'un document et lui assigne une classification multi-étiquettes modélisée par vecteur binaire $\mathbf{y} = (y_1, \ldots, y_k)$. La taille k du vecteur binaire correspond au nombre de mots-clés d'un vocabulaire contrôlé $\mathbb{V}$ dans lequel $y_j = 1$ si le document d est indexé avec le mot-clé contrôlé j et $y_j = -1$ si ce n'est pas le cas.

Formellement, l'opération d'indexation par assignation supervisée peut être représentée de la manière suivante :

$$\mathbb{I}_i^{AS} = \bigcup_{y_j \in \mathcal{G}(\mathbf{d}_i)} \{y_j : y_j = 1\} \qquad [4]$$

La fonction $\mathcal{G}(\mathbf{d}_i)$ est approximée par apprentissage statistique sur une collection d'exemplaires $\{(\mathbf{d}_i, \mathbf{y}_j) \in \mathbb{R}^n \times \{1, -1\}^k\}$. Un document $\mathbf{d}$ peut modéliser plusieurs caractéristiques de son contenu. Généralement, $\mathbf{d}$ correspondra à la distribution des mots (chaînes de caractères et n-grams) qui le composent. L'apprentissage statistique consiste dans ce contexte à induire quelles distributions de mots dans un document sont de bons prédicteurs des mots-clés contrôlés que des documentalistes professionnels sélectionneraient pour l'indexation du document.

Ce sont essentiellement les mêmes algorithmes d'apprentissage qui sont utilisés tant dans les méthodes d'assignation supervisées que dans les méthodes d'extraction supervisées (séparateurs à vaste marge, k plus proches voisins, arbres de décisions, régressions logistiques, modèles bayésiens [HUNL11, JMDA12, MEFÜ08, PYWZ16, PBMH07, TARN09, TPLD09, TLMV13, VACO10, ZPYX15].

2. La campagne d'évaluation DEFT 2016 : l'indexation par extraction vs l'indexation par assignation de mots-clés

Malgré l'importance des tâches d'assignation dans la gestion documentaire, la très grande majorité des contributions scientifiques sur l'automatisation de l'indexation ont porté sur des méthodes d'extraction. À titre d'illustration, parmi les 1 007 publications scientifiques répertoriées dans la figure 1 en introduction, seulement 32 portaient sur des méthodes d'assignation de mots-clés.[3]

Plusieurs raisons peuvent expliquer cette prévalence des méthodes d'extraction par rapport aux méthodes d'assignation. Traditionnellement, selon Lancaster, une partie de l'explication est attribuée à la complexité de la tâche d'assignation des mots-clés :

« The extraction of words and/or phrases from documents is a task that computers can accomplish rather well. […] However, most human indexing is not extraction indexing but assignment indexing and performing this task by computer is altogether more difficult. » [LANC98, p. 256]

À la lumière des travaux des 20 dernières années dans le domaine, nous devons donner raison à Lancaster. Les méthodes automatiques d'extraction ont atteint une certaine maturité en termes de qualité d'indexation que n'ont pas encore atteinte les méthodes d'assignation. Leurs performances ont en effet atteint un niveau fort acceptable dans certains contextes bien balisés. Ainsi, lors de la campagne d'évaluation DEFT 2012, une compétition internationale sur les méthodes d'indexation

[3] C'est-à-dire que seulement 32 références étaient associées à la requête : (TITLE-ABS-KEY("keyphrase assignment") or TITLE-ABS-KEY ("keyword assignment") or TITLE-ABS-KEY("keyterm assignment")).

automatique par extraction, la performance de la méthode gagnante fut caractérisée par une F-mesure (micro-moyenne)[4] de 94.88% [PZFG12]. De plus, on connaît bien maintenant les principales sources d'erreurs générées par les méthodes d'indexation automatique par extraction [HANG14], ce qui n'est pas le cas pour les méthodes d'assignation.

Il en est autrement pour les méthodes d'assignation. La qualité des indexations prédites par des méthodes automatiques d'assignation de mots-clés est généralement beaucoup plus faible que celles réalisées par des méthodes d'extraction automatique de mots-clés. À notre connaissance, la meilleure performance enregistrée lors d'une compétition scientifique portant sur l'évaluation des méthodes d'indexation automatique par assignation a été obtenue au BIOASQ 2015, avec une F-mesure (micro-moyenne) de 63.23% [PYWZ16, ZPYX15]. Toutefois, il est probable que ce résultat soit bien spécifique aux paramètres de la compétition BIOASQ et qu'il soit difficilement généralisable sur d'autres corpus.

En somme, les résultats enregistrés lors de différentes compétitions scientifiques d'indexation automatique semblent corroborer ce qu'affirmait Lancaster, à savoir que l'indexation automatique par assignation est une tâche beaucoup plus difficile que l'extraction. La compétition DEFT-2016 est une occasion de contribuer à cette problématique particulière. En effet, compte tenu des caractéristiques des corpus de la compétition, nous verrons que le type de méthode nécessaire pour gagner la compétition doit être basé sur de l'assignation automatique de mots-clés plutôt que sur de l'extraction.

2.1. *Les caractéristiques des corpus de la compétition*

Le DEFT-2016 suit un protocole relativement classique pour ce genre de compétition scientifique. On demande aux participants de développer une méthode d'indexation automatique de documents qui est ensuite évaluée sur quatre corpus de langue française appartenant à quatre domaines de spécialité scientifique. Le tableau 3 présente les principales caractéristiques de ces quatre corpus. Un premier corpus contient 782 notices liées au domaine de la chimie (CHIMIE), un deuxième corpus contient 706 notices liées au domaine des sciences de l'information (INFO), un troisième corpus contient 715 notices liées au domaine de la linguistique (LING) et un quatrième corpus contient 718 notices liées au domaine de l'archéologie (ARCHEO).

De plus, chaque corpus est associé à un thésaurus du domaine construit par l'INIST-CNRS. Le thésaurus du domaine de la chimie est composé de 40 266 mots-clés contrôlés, le thésaurus du domaine des sciences de l'information est composé de 92 472 mots-clés contrôlés, le thésaurus du domaine de la linguistique contient 13 968 mots-clés contrôlés et le thésaurus du domaine de l'archéologie contient 4 905 mots-clés contrôlés.

Chaque notice est composée d'un titre, d'un résumé et d'une indexation de référence contenant un nombre variable de mots-clés (environ une dizaine) attribués par des documentalistes professionnels. L'évaluation de la méthode d'indexation développée porte sur la capacité de celle-ci à prédire correctement les indexations de référence réalisées par les documentalistes professionnels.

[4] La F-Mesure est une mesure classique en recherche d'information. Elle est basée sur les mesures de rappel et de précision.

	Corpus				
	LING	INFO	ARCHEO	CHIMIE	Moyenne
Nombre de notices	715	706	718	782	730.25
Nombre moyen de mots par document (titre et résumé)	140.69	108.27	199.07	95.65	135.92
Nombre moyen de mots-clés par notice	8.66	8.51	16.55	12.69	11.60
% de mots-clés contrôlés	92.51	87.9	80.59	85.32	86.58
% de mots-clés non-contrôlés	7.5	12.09	19.4	14.69	13.42
% de mots-clés présents	30.71	23.54	38.06	18.12	27.61
% de mots-clés absents	69.3	76.45	61.93	81.89	72.39
% de mots-clés contrôlés présents	27.85	20.53	32.75	17.10	24.56
% de mots-clés contrôlés absents	64.66	67.37	47.84	68.22	62.02
% de mots-clés non-contrôlés présents	2.86	3.01	5.31	1.02	3.05
% de mots-clés non-contrôlés absents	4.64	9.08	14.09	13.67	10.37

Tableau 3. *Caractéristiques des corpus du DEFT 2016.*

Le tableau 3 montre que certains types de mots-clés dominent les corpus. Les notices sont majoritairement composées de mots-clés contrôlés. En moyenne, 86.58% des mots-clés des corpus sont de type contrôlé. De plus, les notices sont majoritairement composées de mots-clés absents du contenu textuel des documents. En effet, en moyenne, 72.39% des mots-clés sont du type absent. Finalement, les mots-clés contrôlés sont aussi principalement des mots-clés de type absent. En moyenne, 62.02% des mots-clés sont à la fois contrôlés et absents des documents.

À la lumière de ces chiffres, on constate dans un premier temps que la tâche de la compétition DEFT-2016 est avant tout une tâche d'indexation par assignation de mots-clés contrôlés. En effet, une méthode qui se limiterait uniquement à une stratégie d'extraction de mots-clés ne pourrait espérer atteindre des performances prédictives qui dépassent le seuil d'environ 27.61%, soit la proportion moyenne de mots-clés présents dans le contenu des documents. De plus, les chiffres du tableau 3 indiquent également que la tâche du DEFT-2016 nécessite une méthode d'assignation basée sur des algorithmes supervisés. Une méthode d'assignation qui mobiliserait uniquement des algorithmes non-supervisés ne pourrait espérer obtenir des performances prédictives de plus de 24.56% environ, soit la proportion moyenne de mots-clés contrôlés présents dans le contenu des documents.

Enfin, le tableau 3 montre aussi que pour la tâche d'indexation du DEFT-2016, une méthode d'indexation qui ne repose que sur l'assignation de mots-clés contrôlés a aussi ses limites. Ses

performances prédictives peuvent au mieux plafonner environ à 86.58%, soit la proportion moyenne de mots-clés contrôlés dans les corpus. Suite à ces observations, pour notre participation à la compétition DEFT-2016, notre équipe a choisi de développer une méthode d'indexation par assignation supervisée[5].

2.2. *La méthode développée*

L'hypothèse à la base de notre méthode est simple. Elle consiste à conjecturer des dépendances statistiques entre d'une part des distributions de mots dans les documents d'un corpus d'apprentissage et d'autre part des mots-clés d'indexations de référence. Autrement dit, notre méthode est basée sur l'hypothèse que l'assignation par un documentaliste d'un mot-clé à un document est corrélée à la présence de certaines formes (simples ou complexes) dans le document (i.e. mots) et cette corrélation serait suffisamment forte pour permettre de prédire les indexations de références attribuées au document. Par exemple, si nous observons régulièrement dans un corpus d'apprentissage que, lorsqu'un document contient la combinaison des mots « ring », « uppercut » et « jab », il est généralement indexé avec le mot-clé « Boxe », nous pouvons conjecturer qu'il y a une dépendance entre la combinaison (ring, uppercut, jab) et « Boxe ». Il n'est pas nécessaire pour une méthode prédictive de connaître la nature linguistique de cette dépendance statistique. Ces dépendances peuvent instancier plusieurs phénomènes linguistiques, cognitifs et même sociaux qui caractérisent l'indexation d'un document. Par exemple, la dépendance peut correspondre à une relation de synonymie, ou encore instancier une hyperonymie. La corrélation entre une distribution de formes et un mot-clé peut refléter des phénomènes très complexes comme les représentations sociales ou les biais cognitifs des documentalistes. Pour notre méthode, ceci est secondaire[6] ; l'important est que ces relations sémantiques, biais, représentations sociales se manifestent empiriquement dans un corpus par des régularités empiriques que nous pouvons exploiter.

La méthode que nous avons développée pour la compétition DEFT 2016 cherche à modéliser et exploiter ces dépendances statistiques entre contenu d'un document et mots-clés d'indexation. Elle procède par la construction d'un espace sémantique des mots-clés d'un vocabulaire contrôlé et par la mobilisation d'un algorithme d'apprentissage supervisé de type k plus proches voisins.

Une manière intuitive d'introduire notre méthode est à l'aide d'une représentation graphique. Ainsi, la figure 2 illustre un sous-espace de l'espace sémantique des mots-clés contrôlés du thésaurus du domaine des sciences de l'information. Dans cet espace, chaque point représente un mot-clé du thésaurus, la proximité spatiale entre deux mots-clés représente plus ou moins ce que nous pouvons appeler leur similarité sémantique, le triangle représente un document test et le rayon formé autour du triangle représente l'indexation prédite par notre méthode. Prédire l'indexation d'un document consiste à lui attribuer les k mots-clés les plus proches dans l'espace sémantique.

[5] Il aurait été envisageable de développer une méthode hybride d'indexation par assignation et par extraction, mais cette approche est très complexe à réaliser et peut introduire une quantité importante de bruit dans les résultats. C'est pour cela que nous ne l'avons pas retenue dans le cadre de ce projet.

[6] Qui plus est, très peu d'études se sont penchées sur l'explication des facteurs déterminants les pratiques d'indexation des documentalistes professionnels [ANPÉ01, MAI99].

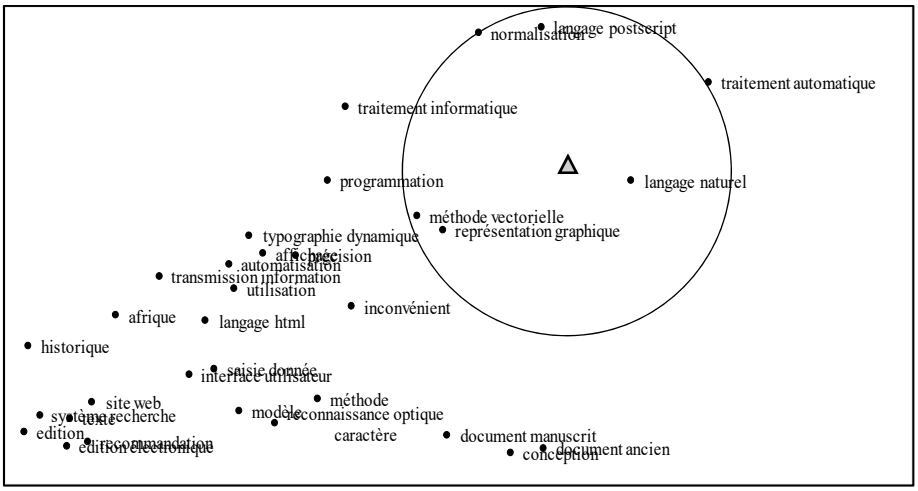

Figure 2. *Illustration d'un sous-espace vectoriel des mots-clés du corpus d'apprentissage INFO*

Dans notre exemple, le document test est indexé avec les mots-clés « langage naturel », « méthode vectorielle », « représentation graphique » et « traitement automatique ». Pour des lecteurs qui connaissent un peu le domaine du TALN, cette combinaison de mots-clés n'est pas surprenante. Ils expriment des thèmes fréquemment associés dans cette littérature. Ceci illustre bien à quoi réfère la notion de « similarité sémantique » modélisée dans un espace sémantique de mots-clés : les mots-clés sont proches dans l'espace lorsqu'ils sont régulièrement assignés conjointement à des documents partageant un contenu similaire. Par conséquent, ils sont corrélés à des distributions de mots similaires dans le corpus d'apprentissage.

Une présentation plus détaillée de la méthode est faite dans la section suivante. La chaîne de traitement est divisée en trois principales étapes, soit le prétraitement des documents, la construction de l'espace sémantique des mots-clés et la prédiction des mots-clés. Elle implique également l'opérationnalisation de deux principaux paramètres, soit 1) celle du coefficient d'association entre les mots-clés et les formes présentes dans les documents et 2) celle de la métrique de proximité entre les documents et les mots-clés.

2.2.1. *Prétraitement des données*

La première étape regroupe plusieurs opérations classiques de prétraitement des documents, notamment de filtrage et de normalisation des occurrences, qui a pour but de construire un dictionnaire de mots pour chacun des corpus. Toutes les occurrences correspondant à des nombres, à des singletons (occurrence d'un seul caractère) ou à des mots fonctionnels (articles, pronoms, déterminant, etc.) ont été filtrées des corpus. Les mots fonctionnels, compte tenu de leur indétermination sémantique, n'entretiennent aucune dépendance statistique avec les mots-clés d'un thésaurus et peuvent donc être filtrés sans affecter les performances prédictives de la méthode. Les occurrences sont ensuite normalisées par racinisation [PORT80], une technique classique qui permet de réduire les flexions d'un même mot. Toutes les séquences de deux mots et plus qui sont répétées plus de deux fois dans un document ont également été récupérées pour représenter le contenu des documents. Cette analyse des segments répétés est une heuristique simple et efficace pour identifier des syntagmes importants dans un document [LESA94].

Les corpus ont finalement été séparés afin de constituer un jeu de données pour l'apprentissage du modèle (composé des deux tiers du corpus) et un jeu de test pour l'évaluation (composé du tiers restant).

2.2.2. *Construction de l'espace sémantique*

La deuxième étape de la méthode est l'apprentissage statistique d'un modèle prédictif. Le modèle construit pour la compétition est inspiré des travaux sur la sémantique vectorielle [BULE07, KICL14, RIEG92, SAHL06, WIDD04]. Formellement, le modèle construit est une matrice dénotée $\left[c(t_i, w_j)\right]^{T\times M} \in \mathbb{R}^M$, où T est le nombre de mots-clés différents présents dans les documents du jeu d'apprentissage et M est le nombre de formes différentes présentes dans les documents après prétraitement. La valeur de $c(t_i, w_j)$ correspond à un coefficient d'association entre un mot-clé d'indexation t_i et une forme w_j. Chaque vecteur $\mathbf{t}_\mathrm{i} = \left(c(t_i, w_1), \ldots c(t_i, w_M)\right)$ de la matrice modélise le patron d'associations qu'un mot-clé entretient avec les formes d'un corpus. Ces vecteurs sont basés sur notre hypothèse de départ, ils modélisent les dépendances statistiques entre mots-clés et les distributions des formes observables dans un corpus d'apprentissage.

L'opérationnalisation de ce coefficient d'association est une étape centrale de la méthode. Plusieurs alternatives sont possibles et certaines d'entre elles ont été comparées lors de nos analyses (voir section 3.1).

Ces coefficients permettent tous de calculer la dépendance statistique entre t_i et w_j dans un corpus donné. Plus t_i est spécifique à w_j, plus la valeur du coefficient est élevée. À notre connaissance, les premières expérimentations en indexation automatique par assignation mobilisant ce type de coefficient furent les travaux de Plaunt et Norgard basés sur l'utilisation du ratio de vraisemblance (G-test) [PLNO98]. D'autres coefficients ont été largement étudiés dans le domaine du traitement automatique des langues, notamment sur la modélisation des phénomènes de colocation [MASC99, Chapitre 5]. La démarche de Plaunt et Norgard, bien qu'elle ne fût pas basée sur des espaces vectoriels de mots-clés, a démontré que les dépendances statistiques entre mots-clés et distribution de formes dans un document étaient suffisamment fortes pour servir de prédicteurs dans un modèle d'assignation. Durant nos expérimentations, nous avons comparé l'impact de ces différents coefficients sur la qualité des prédictions de notre modèle et avons conclu que le coefficient du χ^2 était supérieur aux autres.

2.2.3. *La prédiction des mots-clés*

La troisième étape de la méthode est la prédiction des mots-clés d'indexation pour les documents du jeu de test. Cette étape est basée sur la technique des k plus proches voisins. C'est une technique qui consiste à prédire une variable dépendante à l'aide d'un calcul métrique. Dans le cadre de notre approche, cela consiste à prédire à partir du contenu d'un document les k mots-clés d'indexation d'un thésaurus les plus proches. Soit $\mathbf{d}_j = \left(w_{j1} \ldots w_{jM}\right)$ un vecteur binaire qui représente la distribution des formes présentes dans un document et $\mathbb{T}$ un thésaurus, prédire l'indexation $\mathbb{I}_j = \{t_1, \ldots, t_k\} \in \mathbb{T}$ d'un document j consiste à maximiser la fonction suivante :

$$\mathbb{I}_j = \{t_1, \ldots, t_k\} = \mathrm{argmax}_{t_i \in \mathbb{T}} \sum_{i=1}^{k} sim\left(\mathbf{t}_i, \mathbf{d}_j\right) \qquad [5]$$

Plusieurs métriques de similarité peuvent être implémentées dans une méthode des k plus proches voisins. Nous avons analysé l'impact de plusieurs d'entre elles dans la section 3.1. Ces expérimentations suggèrent que la métrique angulaire du cosinus qui celle qui permet d'optimiser la qualité des indexations prédites par notre méthode.

Le paramètre k, c'est-à-dire le nombre de mots-clés prédits pour chaque document doit aussi être estimé. Nos expérimentations ont montré que le nombre de mots-clés de référence dans les corpus d'apprentissage varie peu : les documents d'apprentissage du corpus INFO ont en moyenne 7.74 mots-clés avec un écart-type σ=3.20, les documents d'apprentissage du corpus CHIMIE ont en moyenne 12.31 mots-clés avec σ=5.18, ceux du corpus LING ont en moyenne 8.76 mots-clés avec σ= 2.11 et les documents d'apprentissage du corpus ARCHEO ont 16.24 mots-clés en moyenne et σ=6.56. Par

conséquent, estimer k à l'aide la moyenne empirique observée dans nos corpus d'apprentissage s'est avérée une approximation très satisfaisante.

2.3. *Résultats des évaluations*

L'évaluation des méthodes développées par les participants du DEFT 2016 est réalisée à l'aide des macro-moyennes de précision, de rappel et de F-mesure sur les quatre corpus. Ce sont des coefficients classiques d'évaluation des faux positifs (bruits) et des faux négatifs (silence) d'une prédiction.

Huit équipes ont participé à la compétition DEFT-2016 et cinq d'entre elles, incluant nous-mêmes, ont soumise leurs méthodes à la campagne d'évaluation [DAIL16], soit l'équipe du LIMSI (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur) [HAMO16], LINA (Laboratoire d'Informatique de Nantes Atlantique) [BOBD16], le LIPN (Laboratoire d'Informatique de Paris Nord) [BUZA16], EXENSA (la société SAS eXenSa) [MAFP16] et notre équipe l'EBSI (École de Bibliothéconomie et des Sciences de l'Information) [CHFL16]. Les résultats détaillés de l'évaluation des méthodes sont présentés dans le tableau 6 et les résultats globaux sur l'ensemble des quatre corpus sont présentés dans la figure 3.

Corpus	Participants	Macro F-mesure	Macro précision	Macro rappel
INFO	EBSI	27.94	31.78	25.81
	EXENSA	23.86	21.26	30.32
	LINA	21.45	21.93	21.83
	LIMSI	12.49	13.83	12.01
	LIPN	15.34	11.72	23.54
LING	EBSI	33.49	32.56	35.33
	EXENSA	26.30	23.28	32.73
	LINA	24.19	23.16	25.85
	LIMSI	15.63	15.67	16.10
	LIPN	19.07	13.98	30.81
CHIMIE	EBSI	21.53	22.94	22.28
	EXENSA	21.46	24.92	21.73
	LINA	18.28	21.15	17.54
	LIMSI	15.29	18.19	14.90
	LIPN	15.31	10.88	30.25
ARCHEO	EBSI	39.43	40.87	40.37
	EXENSA	45.59	43.48	52.71
	LINA	40.11	53.77	33.46
	LIMSI	43.26	55.26	38.03
	LIPN	30.75	33.93	31.25

Tableau 4. *Résultats des évaluations des méthodes des participants au DEFT 2016.*

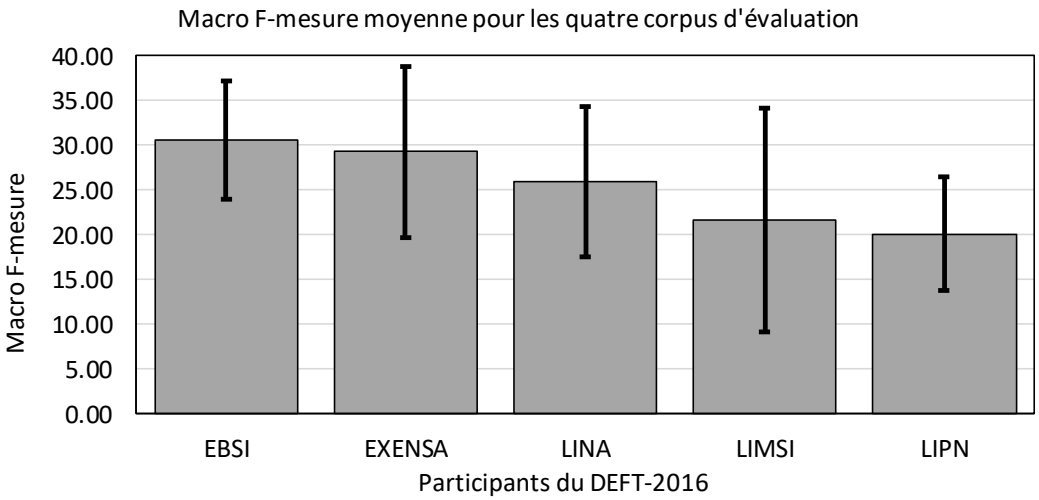

Figure 3. *Comparaison de la macro F-mesure moyenne (MFMM) sur les quatre corpus d'évaluation INFO, LING, CHIMIE et ARCHEO, avec écart-type.*

Les résultats des évaluations montrent que notre méthode s'est classée parmi les meilleures de la compétition. Sur les quatre tâches d'indexation de la compétition, notre méthode s'est classée première trois fois (sur les corpus INFO, LING et CHIMIE) et quatrième sur la tâche d'indexation du corpus ARCHEO.[7] Sur l'ensemble des corpus, c'est notre méthode qui a prédit les meilleures indexations avec une macro F-mesure moyenne (MFMM) de 30.59. Elle est suivie de près par la méthode développée par EXENSA caractérisée par MFMM=29.30. Ensuite c'est la méthode de LINA qui a donné les meilleurs résultats avec MFMM=26.01, la méthode de LIMSI avec MFMM=21.67 et finalement la méthode de LIPPN avec MFMM=20.12.

C'est sur le corpus ARCHEO que notre méthode s'est le moins bien classée comparativement aux autres participants, bien que ce soit pourtant sur ce corpus que notre méthode a obtenu la meilleure F-mesure. D'une part, ceci suggère que le corpus ARCHEO est différent des trois autres corpus. L'analyse du tableau 3 résumant les caractéristiques des quatre corpus corrobore la première observation. En effet, le corpus ARCHEO se différencie des trois autres corpus au niveau du type de mot-clé utilisé dans les indexations de référence. Contrairement aux autres corpus, le corpus ARCHEO est caractérisé par la plus petite proportion de mots-clés contrôlés, la plus grande proportion de mots-clés présents dans les documents et la plus grande proportion de mots-clés contrôlés présents dans les documents.

D'autre part, ces résultats suggèrent également qu'il y a une différence importante dans le type de méthode développée par les autres équipes qui ont participé au DEFT-2016. Nous discutons plus en détail ces différences dans la section 3.2.

Une autre caractéristique intéressante illustrée par les résultats de la figure 3 est la robustesse de notre méthode. Notre méthode est l'une des plus stables parmi les participants, c'est-à-dire que la qualité des indexations prédites varie peu selon les différents corpus. La valeur de MFMM est caractérisée par un écart-type de seulement σ=6.62, comparativement à σ=9.56 pour EXENSA, σ=12.53 pour LIMSI et σ=8.41 pour LINA. La méthode de LIPN est la plus robuste avec σ=6.33, mais elle est également celle caractérisée par une MFMM la plus faible. Ceci constitue un indicateur encourageant concernant la capacité de notre méthode à être généralisée éventuellement sur d'autres corpus que ceux de la compétition. Non seulement notre méthode donne de très bons résultats, ceux-ci

[7] L'évaluation de notre méthode est un peu différente de l'évaluation officielle de la compétition. Dans les résultats officiels, nous avons terminé deuxième sur la tâche d'indexation du corpus CHIMIE, avec un F-mesure de 21.07 plutôt que 21.53. Cette différence est causée par la correction d'une erreur dans la version précédente de notre algorithme. Nous avons choisi d'afficher les résultats de notre propre évaluation car c'est sur ceux-ci que porteront nos analyses dans la quatrième section de l'article.

sont également très stables. Nous conjecturons que cela s'explique par la simplicité de notre approche, qui n'implique de calibrer que très peu de paramètres, deux en particulier, soit un coefficient d'association et une métrique de similarité.

Finalement, notre méthode est la mieux balancée en termes de précision et de rappel. L'écart moyen (RMSE – root mean square error) entre le rappel et la précision de notre méthode est de seulement 6.63, comparativement à 16.33 pour EXENSA, 20.80 pour LINA, 17.64 pour LIMSI et 28.37 pour LIPN. Autrement dit, contrairement aux méthodes de LIPN et d'EXENSA, qui ont tendance à sacrifier la précision de leurs prédictions pour leur rappel, et contrairement aux méthodes de l'IMSI et de LINA (dans une moindre mesure), qui ont tendance à faire l'inverse, notre méthode n'est pas biaisée vers un type d'erreur particulier (bruit vs silence).

3. Analyses

Les résultats présentés dans la section précédente sont prometteurs. Ils suggèrent qu'une approche relativement simple, basée sur un modèle d'espaces sémantiques de mots-clés, constitue une solution efficace permettant d'automatiser des tâches d'indexation par assignation de mots-clés.

Trois types d'analyse ont été effectués sur ces résultats. La première porte sur les paramètres de la méthode. L'objectif est d'évaluer si l'opérationnalisation de notre méthode est optimale ou si d'autres choix d'opérationnalisation sont meilleurs. La deuxième analyse consiste à comparer notre méthode avec celles des autres équipes participantes au DEFT-2016. L'objectif est d'apporter des éléments de réponse expliquant les écarts de performance entre ces méthodes et la nôtre. Finalement, la troisième analyse porte sur la courbe d'apprentissage de notre méthode. L'objectif consiste à expliquer les variations de performances sur les documents des corpus obtenues par notre méthode.

3.1. *Analyse des paramètres de la méthode*

La méthode développée est basée sur l'opérationnalisation de deux principaux paramètres, soit un coefficient d'association et une métrique. L'opérationnalisation utilisée pour la compétition DEFT-2016 est basée sur le chi-carré et la métrique du cosinus. Afin de vérifier si cette opérationnalisation est optimale, nous l'avons comparé avec quatre coefficients d'association alternatifs et quatre autres métriques.

Les coefficients d'association alternatifs avec lesquels le chi-carré a été comparé sont le gain d'information, le ratio de vraisemblance, l'information mutuelle (PMI) et la corrélation de Matthews. Ce sont des coefficients couramment utilisés en TALN [MASC99]. Des définitions basées sur une table de contingence sont présentées dans le tableau 4, dans lesquelles n_{11} est le nombre de documents du jeu d'apprentissage indexés avec le mot-clé t_i et contenant (dans le résumé ou le titre) une occurrence de la forme w_j, n_{01} est le nombre de documents du jeu d'apprentissage non indexés avec le mot-clé t_i et contenant une occurrence de la forme w_j, n_{10} est le nombre de documents du jeu d'apprentissage indexés avec t_i mais ne contenant pas w_j, n_{00} est le nombre de documents d'apprentissage non indexés avec t_i et ne contenant pas w_j et N est le nombre de documents du corpus d'apprentissage.

Gain d'information (GI)	$c(t_i, w_j) = \sum_{i=1}^{2}\sum_{j=1}^{2}\left(\frac{n_{ij}}{N}\log_2\frac{N \times n_{ij}}{(n_{i1}+n_{i0}) \times (n_{1i}+n_{0j})}\right)$
Ratio de vraisemblance (G-test)	$c(t_i, w_j) = 2\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}\left[\log_e(n_{ij}) - \log_e\left(\frac{(n_{i1}+n_{i0}) \times (n_{1i}+n_{0j})}{N}\right)\right]$
Information mutuelle (pmi)	$c(t_i, w_j) = \log_2\left(\frac{n11/N}{\frac{n11+n10}{N} \times \frac{n11+n01}{N}}\right)$
Chi-carré (x^2)	$c(t_i, w_j) = \frac{N((n_{11} \times n_{00}) - (n_{01} \times n_{10}))^2}{(n_{11}+n_{01})(n_{10}+n_{00})(n_{11}+n_{10})(n_{01}+n_{00})}$
Matthews (MCC)	$c(t_i, w_j) = \frac{(n_{11}n_{00}) - (n_{10}n_{01})}{\sqrt{(n_{11}+n_{10})(n_{11}+n_{01})(n_{00}+n_{10})(n_{00}+n_{01})}}$

Tableau 5. *Définitions de plusieurs coefficients d'association*

Le graphique de la figure 5 montre les performances de notre méthode lorsqu'elle est opérationnalisée à l'aide des coefficients d'association alternatifs présentés plus haut. Selon la macro F-mesure moyenne, notre méthode basée sur le chi-carré est l'opérationnalisation qui donne les meilleurs résultats. Les coefficients du gain d'information, le ratio de vraisemblance et la corrélation de Matthews sont également des alternatives intéressantes, alors que l'information mutuelle fait chuter les performances de la méthode. Cette dernière observation est surprenante lorsque l'on sait que l'information mutuelle l'un des coefficients les plus utilisés pour construire des espaces vectoriels de mots [BULE07, KICL14]. Nous verrons plus loin que l'usage de ce coefficient a probablement contribué aux résultats mitigées qu'ont obtenus certaines équipes à la compétition du DEFT-2016.

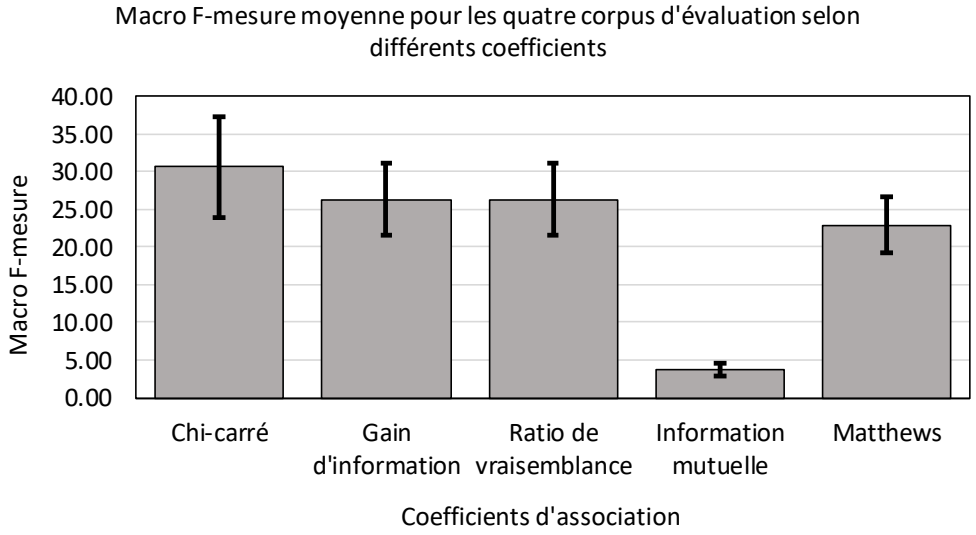

Figure 4. *Comparaison de l'impact de différents coefficients d'association dans les performances de la méthode développée*

Le tableau 5 présente les métriques que nous avons comparées, soit la métrique L2 (inversée), la corrélation de Pearson, le produit scalaire, l'indice de Jaccard et le cosinus. Rappelons qu'un vecteur $\mathbf{t}_i = \left(c(t_i, w_1), \ldots c(t_i, w_M)\right)$ modélise le patron d'associations qu'un mot-clé t_i entretient avec les

formes $w_1 \dots w_M$ d'un corpus et que $\mathbf{d}_j = (w_{j1} \dots w_{jM})$ est un vecteur binaire qui représente la distribution des formes présentes dans un document.

L2	$sim(\mathbf{t}_i, \mathbf{d}_j) = \sqrt{\sum_{k=1}^{m} (c(t_i, w_k) - w_{jk})^2}^{-1}$
Cosinus	$sim(\mathbf{t}_i, \mathbf{d}_j) = \frac{\mathbf{t}_i \cdot \mathbf{d}_j}{\lvert\mathbf{t}_i\rvert \cdot \lvert\mathbf{d}_j\rvert}$
Produit scalaire	$sim(\mathbf{t}_i, \mathbf{d}_j) = \mathbf{t}_i \cdot \mathbf{d}_j$
Jaccard	$sim(\mathbf{t}_i, \mathbf{d}_j) = \frac{\sum_{k=1}^{m} \min(c(t_i, w_k), w_{jk})}{\sum_{k=1}^{m} \max(c(t_i, w_k), w_{jk})}$
Pearson	$sim(\mathbf{t}_i, \mathbf{d}_j) = \frac{\sum_{k=1}^{m} (c(t_i, w_k) - \bar{c}_i)(w_{jk} - \bar{w}_j)}{\sqrt{\sum_{k=1}^{m} (c(t_i, w_k) - \bar{c}_i)^2} \sqrt{\sum_{i=1}^{m} (w_{jk} - \bar{w}_j)^2}}$

Tableau 6. *Métriques de similarité comparées lors de nos expérimentations*

Nos expérimentations suggèrent que c'est la métrique angulaire du cosinus qui permet d'optimiser la qualité des indexations prédites par notre méthode.

Le graphique de la figure 6 montre les performances de notre méthode lorsqu'elle est opérationnalisée à l'aide des métriques alternatives présentées plus haut. Selon la macro F-mesure moyenne, la métrique angulaire du cosinus est l'opérationnalisation qui permet d'optimiser la qualité des indexations prédites par notre méthode, suivie de très près par la corrélation de Pearson. Ensuite, les meilleurs résultats sont obtenus à l'aide de l'indice de Jaccard et le produit scalaire. La métrique L2 est l'opérationnalisation qui fait chuter le plus les performances de notre méthode. Ces résultats d'analyse montrent que le choix de la métrique dans ce type de méthode est déterminant.

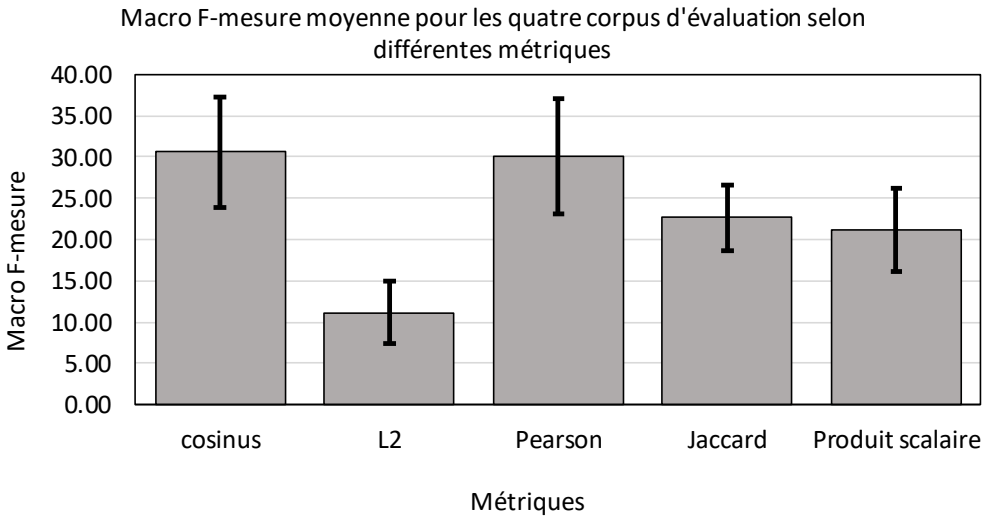

Figure 5. *Comparaison de l'impact de différentes métriques dans le paramétrage de la méthode développée.*

3.2. *Comparaison avec les méthodes des autres équipes participantes*

Quatre autres équipes ont participé à la compétition. Nous ne pouvons pas ici pour des raisons d'espace présenter dans le détail les méthodes développées par les autres participants. Une présentation détaillée se retrouve les actes du DEFT-2016. Nous proposons néanmoins de classer les méthodes développées par les participants selon la typologie des méthodes d'indexation automatique que nous avons proposée dans la section 1.2. Nous proposons ensuite certaines conjectures afin d'expliquer pourquoi ces méthodes performent moins bien que la nôtre en termes de qualité d'indexation produite.

L'équipe LIPN [BUZA16] a développé une méthode d'indexation par assignation supervisée dont le design est très similaire à celui de notre méthode. L'une des principales différences entre notre méthode et la leur est le choix d'opérationnalisation des paramètres. Le LIPN a utilisé l'information mutuelle comme coefficient d'association et le produit scalaire comme métrique. Les résultats obtenus par le LIPN sont beaucoup plus faibles que les nôtres, malgré la similitude de design et comme l'illustrent les résultats obtenus dans la section précédente, ceci est lié aux choix d'opérationnalisation de ces deux paramètres. L'information mutuelle est de loin le coefficient le moins performant de ceux que nous avons comparés et le produit scalaire est également l'une des opérationnalisations les moins performantes que nous avons comparées. Les raisons expliquant les faibles performances associées à ces paramètres sont bien connues. Le produit scalaire est une métrique qui ne tient pas compte de la magnitude des vecteurs. Normer l'espace vectoriel aurait permis d'obtenir de bien meilleurs résultats. C'est d'ailleurs la raison pour laquelle la métrique du cosinus performe aussi bien. En ce qui concerne l'information mutuelle, il s'agit d'un coefficient d'association très sensible aux fréquences faibles ($\leq$5 environ) pour lesquelles il a tendance à surestimer l'association entre les deux variables [BOUM09, MASC99]. Or, la majorité des mots-clés contrôlés des corpus d'apprentissage sont caractérisés par des fréquences de cette magnitude. Les coefficients alternatifs comme le chi-carré et le gain d'information n'ont pas ce biais.

Le LIMSI [HAMO16] a développé une méthode d'indexation par assignation non-supervisée basée sur une analyse terminologique. Dans un premier temps, pour chaque notice, la méthode extrait grâce à l'analyse terminologique un ensemble de mots-clés candidats qu'elle compare dans un deuxième temps avec les mots-clés contrôlés des thésaurus des domaines de spécialité. Les mots-clés candidats qui correspondent (en partie ou en totalité) avec l'un des mots-clés contrôlés d'un thésaurus sont ensuite ordonnés par importance selon leur rang d'apparition dans une notice. Finalement les n mots-clés les plus importants sont assignés à la notice. La méthode performe bien pour le corpus ARCHEO, mais les scores de F-mesure chutent pour les trois autres corpus LING, CHIMIE et INFO. La raison qui explique ces faibles performances est liée au type de méthode utilisée (assignation non-supervisée) par le LIMSI, qui ne permet pas de prédire les mots-clés de référence absents des notices. En effet, les performances de la méthode du LIMSI sont fortement contraintes par l'étape d'extraction terminologique. Or, comme nous l'avons mentionné précédemment, en moyenne, 72.39% des mots-clés de références sont absents des notices et ne peuvent faire l'objet d'une extraction terminologique. Seul le corpus ARCHEO est un peu différent à cet égard. En effet, dans ce corpus, 61.93% des mots-clés de références sont absents. Puisque la majorité des mots-clés de référence sont absents du contenu des notices (titre et résumé), ils doivent être inférés, ce que ne permet pas de faire la méthode développée par le LIMSI.

La méthode développée par l'équipe du LINA [BOBD16] est une méthode d'indexation automatique basée sur l'analyse de graphe à la manière de TextRank [MITA04]. Les participants ont développé trois variantes de leur méthode, mais celle qui a produit les meilleurs résultats est une méthode de type assignation non-supervisée. La méthode suit des étapes similaires à celle du LIMSI, la principale différence est le critère d'ordonnancement, qui est ici basé sur la connexité des mots-clés dans un graphe de cooccurrence. Parce qu'elle est une méthode de type assignation non-supervisée, la méthode du LINA rencontre les mêmes limites que la méthode du LIMSI, c'est-à-dire que seulement les mots-clés contrôlés présents dans les documents peuvent être attribués à une nouvelle notice. C'est ce qui explique les résultats plus faibles obtenus pour les corpus INFO, LING et CHIMIE, dans

lesquels une part plus importante des mots-clés contrôlés sont absents du contenu des notices et doivent par conséquent être inférés.

La méthode développée par l'équipe EXENSA [MAFP16] est une combinaison de deux méthodes complémentaires d'indexation automatique. La première est une approche de type assignation non-supervisée, basée sur la recherche de similarité graphique (via une analyse des n-gram de caractères) entre les mots-clés contrôlés des thésaurus et le lexique des notices de test. Lorsque cette méthode est utilisée seule, elle rencontre les mêmes difficultés que les méthodes développées par le LIMSI et le LINA, c'est-à-dire qu'elle performe beaucoup moins bien sur des corpus comme INFO, LING et CHIMIE pour lesquels les indexations de référence sont majoritairement composées d'une part très importante de mots-clés contrôlés absents du contenu des notices. La deuxième méthode est une approche de type assignation supervisée, basée sur la construction d'une espace vectoriel des notices. Elle consiste à récupérer pour une nouvelle notice test, les k notices d'apprentissage les plus similaires et à attribuer à celle-ci les mots-clés fortement associés à celles-là. L'hypothèse derrière cette approche est que des notices au contenu similaire devraient se voir attribuer des mots-clés également similaires. Bien que cette deuxième méthode devrait permettre (à la différence de la première) d'attribuer des mots-clés absents des notices tests, les résultats obtenus sont bien inférieurs à la première méthode. Il est à noter que cette deuxième méthode utilise, comme c'était le cas pour LIPN, le coefficient de l'information mutuelle, lequel cause des biais importants. Nous conjecturons que c'est l'une des raisons expliquant les résultats obtenus par cette équipe. C'est en combinant les deux méthodes qu'EXENSA a obtenu ses meilleurs résultats. Toutefois, étant donné les faibles performances de leur méthode d'assignation supervisée, ces résultats ont les mêmes caractéristiques que ceux obtenus par l'équipe du LIMSI et du LINA : la qualité des indexations est beaucoup plus faible pour les corpus INFO, LING et CHIMIE pour lesquels une part plus importante des mots-clés contrôlés sont absents du contenu des notices et doivent par conséquent être inférés.

En somme, notre méthode se distingue de celles des autres participants au DEFT par ses performances au niveau de l'indexation de mots-clés contrôlés absents du contenu des notices tests. C'est la raison pour laquelle notre méthode performe beaucoup mieux que celles des autres équipes participantes pour les corpus INFO, LING et CHIMIE. Notre méthode semble être la seule véritablement capable d'inférer des mots-clés absents du contenu des notices. Par design, la méthode du LIPN en a potentiellement la capacité également, mais elle fut basée sur de mauvais choix d'opérationnalisation au niveau du coefficient d'association et de la métrique utilisés, ce qui explique probablement ses moins bonnes performances.

3.3. *Variation des performances et courbe d'apprentissage*

Par ailleurs, les résultats obtenus montrent également que l'indexation par assignation de mots-clés est une tâche très difficile à automatiser et que les solutions actuelles ne sont pas encore pleinement satisfaisantes. En moyenne, les indexations prédites par notre méthode ont une macro F-mesure de 30.59, mais cette moyenne est associée à une variation par document très importante. La qualité des indexations prédites pour certains documents de test peut atteindre une F-mesure de 80.00, alors qu'elle est totalement erronée pour d'autres documents, c'est-à-dire que sa F-mesure est de 0.0. La figure 4 illustre cette variation sous la forme d'un diagramme en boîte.

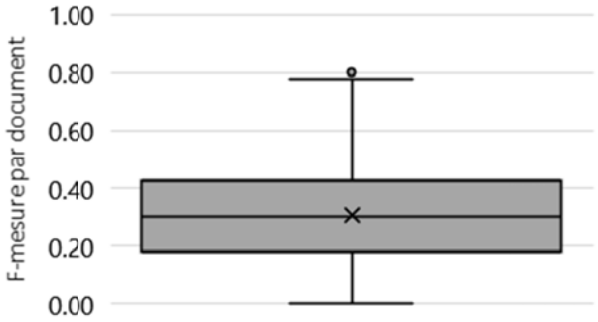

Figure 6. *Variation de la F-mesure par document selon les indexations prédites par notre méthode.*

Comme l'illustre le diagramme, 25% des indexations prédites par notre méthode ont une F-mesure qui varie entre 0.0 et 18.18, 50% des indexations prédites ont une F-mesure qui varient entre 18.18 et 42.42 et le dernier quartile regroupe des indexations dont la F-mesure varie entre 42.42 et 80.00.

Ces variations suggèrent que la prédiction des indexations de référence de certains documents est beaucoup plus difficile que pour d'autres documents. La prédiction de l'assignation de certains mots-clés est très difficile et sujette à beaucoup d'erreurs, mais la prédiction d'autres mots-clés est beaucoup plus univoque. Par exemple, prédire l'assignation d'un mot-clé contrôlé pour un document test est impossible dans le cadre de notre méthode si ce mot-clé n'est pas présent au moins une fois dans le corpus d'apprentissage. Autrement dit, s'il n'y a aucun document de la base d'apprentissage associé au mot-clé, aucune dépendance statistique entre ce mot-clé et le contenu d'un document ne peut être calculée et la prédiction est alors impossible. Or, 22% des mots-clés de référence des documents tests du corpus INFO étaient absents du corpus d'apprentissage, 21% des mots-clés de référence des documents tests du corpus LING étaient également absents, pour le corpus CHIMIE, c'était 36% des mots-clés du corpus test qui étaient absents du corpus d'apprentissage et finalement 16% des mots-clés du corpus test ARCHEO étaient absents. La faible représentativité des corpus d'apprentissage rend un modèle prédictif construit par apprentissage statistique extrêmement problématique. Ce modèle souffrira fort probablement de sur-ajustement (overfitting).

Afin d'apporter des éléments de réponses à ces résultats, nous avons mené une analyse de la courbe d'apprentissage de notre méthode. L'analyse de la courbe d'apprentissage consiste à évaluer les indexations prédites par notre méthode que pour les mots-clés plus fréquents qu'un seuil donné. L'objectif de cette analyse est de vérifier si les performances de la méthode sont dépendantes de la représentativité de la base d'apprentissage. Autrement dit, le but est de vérifier à quel point la fiabilité de la prédiction de l'assignation d'un mot-clé dépend du nombre de documents auxquels ce mot-clé a été attribué dans la base d'apprentissage. La procédure est itérative : dans la première itération, nous évaluons notre méthode que sur les mots-clés présents au moins une fois dans la base d'apprentissage (N_min>0), dans une deuxième itération, nous évaluons notre méthode que sur les mots-clés présents au moins deux fois dans la base d'apprentissage (N_min>1) et ainsi de suite (N_min>k) jusqu'à ce que la qualité des prédictions de la méthode plafonne. Nous nous attendons à ce que la valeur de la F-mesure croisse avec la progression de N_min.

Les résultats sont illustrés dans le graphique de la figure 7. Nous avons analysé la courbe d'apprentissage entre N_min>=0 et N_min>20.

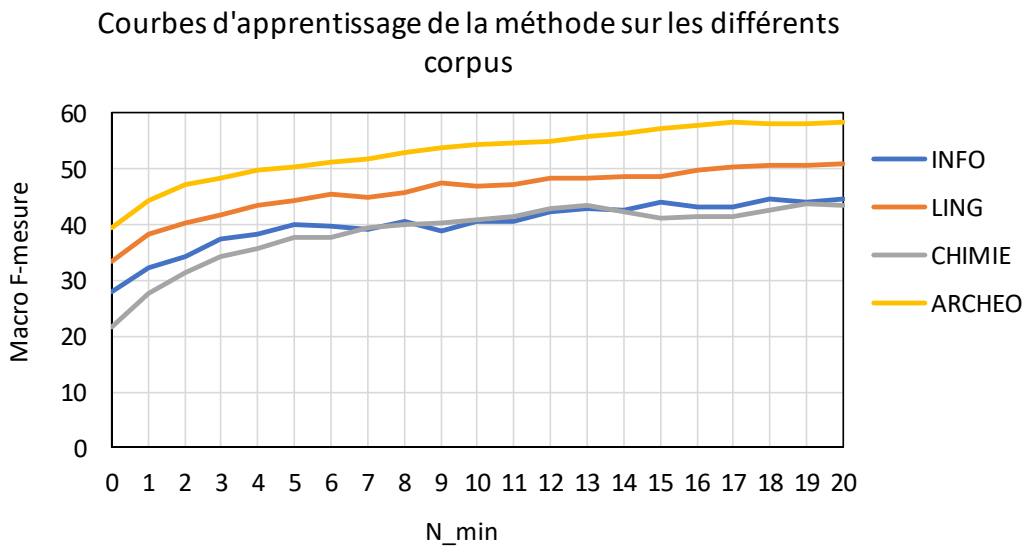

Figure 7. *Courbes d'apprentissages de notre méthode. Évolution de la F-mesure selon la représentativité de la base d'apprentissage, où N_min est égal au nombre minimal de documents d'apprentissage auxquels un mot-clé doit être assigné pour être retenu dans l'évaluation.*

Les résultats de l'analyse de la courbe d'apprentissage de notre méthode corroborent notre hypothèse. Ces résultats montrent clairement que plus la base d'apprentissage est représentative, c'est-à-dire plus les mots-clés sont assignés à un nombre élevé de documents dans la base d'apprentissage, meilleures sont les indexations prédites par notre méthode. Les meilleurs résultats obtenus sont pour le corpus ARCHEO, avec une macro F-mesure de 58.80.

Par ailleurs, la forme de la courbe d'apprentissage est également révélatrice. Elle montre que l'accroissement de la qualité des indexations prédites par notre méthode n'évolue pas de manière linéaire avec N_min. La valeur de la F-mesure augmente très rapide au début et lentement par la suite pour finalement plafonner autour de 20 exemplaires. C'est lorsque 1 <= N_min <= 5 que l'accroissement marginal de la F-mesure est la plus élevée dans les courbes d'apprentissage.

Cette dernière observation peut servir de recommandation pour la constitution future de ressources. En effet, la constitution de corpus indexés par des documentalistes professionnels est très coûteuse (nous avons mentionné en introduction que le coût d'indexation manuelle d'un document par PubMed est estimé à 9,40$ US). Malgré ces coûts, la constitution de ces ressources est indispensable au développement de solutions automatiques d'assignation supervisée de mots-clés, car elles sont nécessaires à l'apprentissage statistique des modèles d'assignation. L'analyse des courbes d'apprentissage indique où se situe l'investissement optimal pour la constitution de ces ressources. Lors de la constitution d'un corpus d'apprentissage, s'assurer que chaque mot-clé est assigné à environ cinq documents représente le meilleur ratio coût/bénéfice.

Conclusion

Cet article rend compte de manière détaillée de notre participation à l'édition 2016 du Défi fouille de textes. Dans un premier temps, nous avons présenté un état de la situation concernant l'importance, mais aussi les enjeux et les défis de l'indexation automatique. À cet égard, nous avons décrit une typologie des grandes familles de mots-clés employés pour indexer les documents. Par la suite, nous nous sommes attardés à décrire les caractéristiques des principales méthodes permettant d'assister l'indexation des documents. Après avoir présenté les grandes lignes de la campagne d'évaluation DEFT 2016, en insistant sur certains traits des corpus employés, nous avons décrit les caractéristiques

de l'approche que nous avons développée. Celle-ci repose sur la construction d'un espace sémantique de mots-clés.

Les analyses que nous avons menées suggèrent que notre méthode est particulièrement adaptée à des tâches d'indexation automatique qui nécessitent une part importante d'assignation de mots-clés contrôlés. Les performances de notre méthode sont plus élevées lorsque les mots-clés à prédire ne sont pas présents directement dans les documents. C'est la raison pour laquelle notre méthode domine la compétition pour les tâches d'indexation des corpus INFO et LING, mais qu'en contrepartie elle est moins performante sur le corpus ARCHEO.

La qualité des indexations automatiques peut à notre avis être encore grandement améliorée. Le développement de stratégies pour l'indexation automatique des documents textuels repose sur une meilleure compréhension des erreurs commises par notre approche et sur le recours à des corpus d'apprentissage plus volumineux. C'est dans cette voie que nous poursuivons nos travaux de recherche.

Références

[ANPÉ01] Anderson, J. D., & Pérez-Carballo, J. (2001). «The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing». Information Processing & Management, 37, 2, p. 231–254.

[BACO00] Barker, K., & Cornacchia, N. (2000). «Using Noun Phrase Heads to Extract Document Keyphrases». In H. J. Hamilton (Éd.), Advances in Artificial Intelligence (p. 40-52). Springer Berlin Heidelberg.

[BASE00] «Bases de données - Institut de l'information scientifique et technique». (s. d.). Consulté 1 février 2017, à l'adresse http://www.inist.fr/?-Bases-de-donnees-&lang=fr

[BEMM15] Beliga, S., Meštrović, A., & Martinčić-Ipšić, S. (2015). «An overview of graph-based keyword extraction methods and approaches». Journal of information and organizational sciences, 39, 1, p. 1–20.

[BOUG15] Bougouin, A. (2015). Indexation automatique par termes-clés en domaines de spécialité (Thèse de doctorat). Université de Nantes, Nantes.

[BOBD16] Bougouin, A., Boudin, F., & Daille, B. (2016). «TopicRank en domaines de spécialité : participation du LINA à DEFT 2016». In JEP-TALN-RECITAL (Vol. 8, p. 41-47). Paris.

[BOUM09] Bouma, G. (2009). «Normalized (pointwise) mutual information in collocation extraction». Proceedings of GSCL, p. 31–40.

[BULE07] Bullinaria, J. A., & Levy, J. P. (2007). «Extracting semantic representations from word co-occurrence statistics: A computational study». Behavior research methods, 39, 3, p. 510-526.

[BUZA16] Buscaldi, D., & Zargayouna, H. (2016). «LIPN@ DEFT2016: Annotation de documents en utilisant l'Information Mutuelle». In JEP-TALN-RECITAL (Vol. 8, p. 27-33). Paris.

[CHFL16] Chartier, J.-F., Forest, D., & Lacombe, O. (2016). «Alignement de deux espaces sémantiques à des fins d'indexation automatique». In JEP-TALN-RECITAL (Vol. 8, p. 13-19). Paris.

[DAIL16] Daille, B. B., Sabine Boudin, Florian Bougoin, Adrien Cram, Damien Hazem, Amir. (2016). «Indexation d'articles scientifiques Présentation et résultats du défi fouille de textes DEFT 2016» (Vol. 8, p. 1-12). Présenté à JEP-TALN-RECITAL 2016, Association Francophone pour la Communication Parlée (AFCP) et Association pour le Traitement Automatique des Langues (ATALA).

[ERCI07] Ercan, G., & Cicekli, I. (2007). «Using lexical chains for keyword extraction». Information Processing & Management, 43, 6, p. 1705–1714.

[FRQU07] Francis, É., & Quesnel, O. (2007). «Indexation collaborative et folksonomies». Documentaliste-Sciences de l'information, 44, 1, p. 58–63.

[FPWG99] Frank, E., Paynter, G. W., Witten, I. H., Gutwin, C., & Nevill-Manning, C. G. (1999). «Domain-specific keyphrase extraction». In Proceeding of 16th International Joint Conference on Artificial Intelligence (p. 668-673). San Fancisco: Morgan Kaufmann Publishers.

[HAMO16] Hamon, T. (2016). «Indexation automatique de notices bibliographiques à l'aide d'approches d'acquisition terminologique». In JEP-TALN-RECITAL (Vol. 8, p. 20-26). Paris.

[HANG14] Hasan, K. S., & Ng, V. (2014). «Automatic Keyphrase Extraction: A Survey of the State of the Art.». In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (p. 1262–1273). Baltimore, Maryland, USA: ACL.

[HUNL11] Huang, M., Névéol, A., & Lu, Z. (2011). «Recommending MeSH terms for annotating biomedical articles». Journal of the American Medical Informatics Association, 18, 5, p. 660–667.

[JMDA12] Jimeno-Yepes, A., Mork, J. G., Demner-Fushman, D., & Aronson, A. R. (2012). «A one-size-fits-all indexing method does not exist: automatic selection based on meta-learning». Journal of Computing Science and Engineering, 6, 2, p. 151–160.

[JONE72] Jones, K. S. (1972a). «A statistical interpretation of term specificity and its application in retrieval». Journal of documentation, 28, 1, p. 11-21.

[KICL14] Kiela, D., & Clark, S. (2014). «A systematic study of semantic vector space model parameters». In Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) at EACL (p. 21-30).

[KIBK10] Kim, S. N., Baldwin, T., & Kan, M.-Y. (2010). «Evaluating N-gram based evaluation metrics for automatic keyphrase extraction». In Proceedings of the 23rd international conference on computational linguistics (p. 572–580). Association for Computational Linguistics.

[KMKB10] Kim, S. N., Medelyan, O., Kan, M.-Y., & Baldwin, T. (2010). «Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles». In Proceedings of the 5th International Workshop on Semantic Evaluation (p. 21–26). Association for Computational Linguistics.

[LANC98] Lancaster, F. W. (1998). Indexing and Abstracting in Theory and Practice (2e éd.). Champaign, Ill.: Univ. of Illinois Graduate School of Information and Library Science.

[LESA94] Lebart, S., & Salem, A. (1994). Statistique textuelle. Paris: Dunod.

[LPWZ15] Liu, K., Peng, S., Wu, J., Zhai, C., Mamitsuka, H., & Zhu, S. (2015). «MeSHLabeler: improving the accuracy of large-scale MeSH indexing by integrating diverse evidence». Bioinformatics, 31, 12, p. i339–i347.

[MAI99] Mai, J.-E. (1999). «Deconstructing the indexing process». In Advances in Librarianship (p. 269–298). Emerald Group Publishing Limited.

[MASC99] Manning, C., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. Cambridge: MIT Press.

[MAFP16] Marchand, M., Fouquier, G., & Pitel, G. (2016). «Représentation vectorielle de mots pour l'indexation de notices bibliographiques». In JEP-TALN-RECITAL (Vol. 8, p. 34-40). Paris.

[MEWI06] Medelyan, O., & Witten, I. H. (2006). «Thesaurus Based Automatic Keyphrase Indexing». In Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries (p. 296-297). New York, NY, USA: ACM.

[MEDL16] «MEDLINE®/PubMed® Resources Guide». (2016). [List of Links]. Consulté 1 février 2017, à l'adresse https://www.nlm.nih.gov/bsd/pmresources.html#statistics

[MEFÜ08] Mencia, E. L., & Fürnkranz, J. (2008). «Efficient pairwise multilabel classification for large-scale problems in the legal domain». In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (p. 50–65). Springer.

[MITA04] Mihalcea, R., & Tarau, P. (2004). «TextRank: Bringing Order into Texts». In Conference on Empirical Methods in Natural Language Processing (EMNLP), 2004, Barcelona, Spain (p. 404-411). Association for Computational Linguistics.

[NGKA07] Nguyen, T. D., & Kan, M.-Y. (2007). «Keyphrase Extraction in Scientific Publications». In D. H.-L. Goh, T. H. Cao, I. T. Sølvberg, & E. Rasmussen (Éd.), Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers (p. 317-326). Springer Berlin Heidelberg.

[PZFG12] Paroubek, P., Zweigenbaum, P., Forest, D., & Grouin, C. (2012a). «Indexation libre et contrôlée d'articles scientifiques Présentation et résultats du défi fouille de textes DEFT2012». In Actes du huitième DÉfi Fouille de Textes. Grenoble, France.

[PYWZ16] Peng, S., You, R., Wang, H., Zhai, C., Mamitsuka, H., & Zhu, S. (2016). «DeepMeSH: deep semantic representation for improving large-scale MeSH indexing». Bioinformatics, 32, 12, p. i70–i79.

[PBMH07] Pestian, J. P., Brew, C., Matykiewicz, P., Hovermale, D. J., Johnson, N., Cohen, K. B., & Duch, W. (2007). «A shared task involving multi-label classification of clinical free text». In Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing (p. 97–104). Association for Computational Linguistics.

[PLNO98] Plaunt, C., & Norgard, B. A. (1998). «An association-based method for automatic indexing with a controlled vocabulary». Journal of the American Society for Information Science, 49, 10, p. 888–902.

[PORT80] Porter, M. F. (1980). «An algorithm for suffix stripping». Program: electronic library and information systems, 14, 3, p. 130-137.

[RIEG92] Rieger, B. B. (1992). «Fuzzy computational semantics». In Fuzzy Systems. Proceedings of the Japanese-German-Center Symposium, Series (Vol. 3, p. 197–217).

[SAHL06] Sahlgren, M. (2006). The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. Stockholm, Stockholm.

[SAWY75] Salton, G., Wong, A., & Yang, C.-S. (1975a). «A vector space model for automatic indexing». Communications of the ACM, 18, 11, p. 613–620.

[SORO10] Sorower, M. S. (2010). «A literature survey on algorithms for multi-label learning». Oregon State University, Corvallis.

[SPAR74] Sparck Jones, K. (1974). «Automatic indexing». Journal of documentation, 30, 4, p. 393–432.

[SSGS12] Suchecki, K., Salah, A. A. A., Gao, C., & Scharnhorst, A. (2012). «Evolution of wikipedia's category structure». Advances in Complex Systems, 15, supp01, p. 1250068.

[TARN09] Tang, L., Rajan, S., & Narayanan, V. K. (2009). «Large scale multi-label classification via metalabeler». In Proceedings of the 18th international conference on World wide web (p. 211–220). ACM.

[THEC17] «The Citaton Connection - Real Facts - Clarivate Analytics». (2017). Consulté le 1er février 2017, à l'adresse http://wokinfo.com/citationconnection/

[TPLD09] Trieschnigg, D., Pezik, P., Lee, V., De Jong, F., Kraaij, W., & Rebholz-Schuhmann, D. (2009). «MeSH Up: effective MeSH text classification for improved document retrieval». Bioinformatics, 25, 11, p. 1412–1418.

[TBMP15] Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M. R., … others. (2015). «An overview of the bioasq large-scale biomedical semantic indexing and question answering competition». BMC bioinformatics, 16, 1, p. 138.

[TSKA06] Tsoumakas, G., & Katakis, I. (2006). «Multi-label classification: An overview». International Journal of Data Warehousing and Mining, 3, 3.

[TLMV13] Tsoumakas, G., Laliotis, M., Markantonatos, N., & Vlahavas, I. (2013). Large-scale semantic indexing of biomedical publications at bioasq. BioASQ Valencia, Spain.

[TURN00] Turney, P. D. (2000). «Learning algorithms for keyphrase extraction». Information retrieval, 2, 4, p. 303–336.

[VANR79] Van Rijsbergen, C. J. (1979). Information Retrieval. London: Butterworths.

[VACO10] Vasuki, V., & Cohen, T. (2010). «Reflective random indexing for semi-automatic indexing of the biomedical literature». Journal of biomedical informatics, 43, 5, p. 694–700.

[WAXI08] Wan, X., & Xiao, J. (2008a). «CollabRank: Towards a Collaborative Approach to Single-document Keyphrase Extraction». In Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1 (p. 969-976). Stroudsburg, PA, USA: Association for Computational Linguistics.

[WIDD04] Widdows, D. (2004). Geometry and Meaning. Standford: CSLI Publications.

[WPFG99] Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (1999). «KEA: Practical Automatic Keyphrase Extraction». In Proceedings of the 4th ACM Conference on Digital Libraries (p. 254-255). New York, NY, USA: ACM.

[ZPYX15] Zhang, Y., Peng, S., You, R., Xie, Z., Wang, B., & Zhu, S. (2015). «The fudan participation in the 2015 bioasq challenge: Large-scale biomedical semantic indexing and question answering». In CEUR Workshop Proceedings (Vol. 1391). CEUR Workshop Proceedings.

