Indexation d'articles scientifiques Présentation et résultats du défi fouille de textes DEFT 2016

Automatic indexing of scientific papers Presentation and results of DEFT 2016 text mining challenge

Béatrice Daille[1], Sabine Barreaux[2], Adrien Bougouin[1], Florian Boudin[1], Damien Cram[1], Amir Hazem[1]

[1] LS2N- Univ. Nantes, 2 rue de la Houssinière, 44322 Nantes Cedex 3, France
prenom.nom@univ-nantes.fr
[2] INIST–CNRS, 2, allée du Parc de Brabois, 54519 Vandœuvre-lès-Nancy, France
sabine.barreaux@inist.fr

RÉSUMÉ. Cet article décrit la campagne 2016 du défi fouille de textes (DEFT), qui pour sa douzième édition a proposé aux participants de simuler la tâche d'indexation de documents scientifiques réalisée par des documentalistes, experts dans des domaines de spécialité. L'indexation consiste à proposer un ensemble de mots-clés pour une notice bibliographique, en français, de quatre domaines de spécialité (linguistique, sciences de l'information, archéologie et chimie). Cette tâche d'indexation de document scientifique est difficile qu'elle soit réalisée manuellement ou automatiquement. Nous présentons la pratique de l'indexation manuelle et les méthodes état de l'art pour l'indexation automatique ainsi que leurs évaluations. Nous décrivons ensuite les données mises à disposition des participants, le déroulement de la campagne et les résultats obtenus évalués avec les mesures de précision, rappel, et f1-mesure, calculées avec une macro-moyenne.

ABSTRACT. This paper presents the 2016 edition of the DEFT text mining challenge. This edition adresses the keyword-based indexing of scientific papers with the aim of simulating a professional indexer. The corpus is composed of French bibliographic records from four domains : linguistics, information science, archaeology and chemisty. The results have been evaluated in terms of precision, recall and f-measure computed on stemmed texts against a reference manual indexation.

MOTS-CLÉS. indexation automatique, mot-clé, domaines de spécialité, articles scientifiques, français.

KEYWORDS. document indexing, keyphrase, specialized domains, scientific aricles, French.

1 Introduction

L'indexation automatique consiste à identifier un ensemble de mots-clés (e.g. mots, termes, noms propres) qui décrit le contenu d'un document. Les mots-clés peuvent ensuite être utilisés, entre autres, pour faciliter la recherche d'information ou la navigation dans les collections de documents. À l'instar de l'édition 2012 de DEFT (Paroubek *et al.*, 2012), la tâche porte sur l'indexation de documents scientifiques par l'intermédiaire de mots-clés. Alors que l'édition 2012 portait sur l'identification des mots-clés choisis par les auteurs, la tâche de l'édition 2016 concerne l'identification des mots-clés fournis par des documentalistes, des indexeurs professionnels spécialisés dans des domaines. L'indexation par mots-clés fournit un ensemble restreint de mots ou expressions qui représentent ses sujets principaux, explicites ou non (voir la figure 1.1).

Contrairement aux mots-clés d'auteurs, ceux proposés par des indexeurs professionnels sont issus d'une démarche documentaire étudiée pour l'indexation de documents dans le contexte de la recherche d'information. S'appuyant sur le contenu du document et sur un thésaurus du domaine, les indexeurs professionnels fournissent des mots-clés cohérents et exhaustifs. La cohérence implique qu'un concept est toujours représenté par le même mot-clé pour les documents d'un même domaine. Le thésaurus du domaine est donc privilégié pour l'identification

La cause linguistique

L'objectif est de fournir une définition de base du concept linguistique de la cause en observant son expression. Dans un premier temps, l'A. se demande si un tel concept existe en langue. Puis il part des formes de son expression principale et directe (les verbes et les conjonctions de cause) pour caractériser linguistiquement ce qui fonde une telle notion.

Mots-clés de référence : français ; interprétation sémantique ; conjonction ; expression linguistique ; concept linguistique ; relation syntaxique ; cause.

FIGURE *1.1.* *Exemple d'indexation par mots-clés d'une notice bibliographique (résumé). Les mots-clés soulignés sont explicites, c'est-à-dire qu'ils occurrent dans le document, les autres sont implicites.*

des mots-clés. L'emploi d'un référentiel pour indexer des textes est appelée *indexation contrôlée*. Toutefois, l'exhaustivité implique aussi que l'indexeur fournisse des mots-clés relatifs à des concepts importants n'appartenant pas nécessairement au thésaurus. Le non-emploi d'un vocabulaire contrôlé pour indexer les textes est appelé *indexation libre*.

Les méthodes mises au défi dans cette édition 2016 doivent identifier les concepts importants permettant d'indexer les documents. Comme l'indexation proposée par les indexeurs professionnels, les méthodes pourront effectuer une indexation contrôlée, libre ou mixte.

Nous présentons successivement les pratiques de l'indexation manuelle, un état de l'art sur les méthodes de référence en indexation libre et en indexation contrôlée, les données qui ont été mises à la disposition des participants, la tâche proposée, les mesures d'évaluation et les résultats obtenus par les participants.

2 Pratiques de l'indexation manuelle

Les pratiques prises en compte dans ce défi sont celles mises en œuvre à l'INIST-CNRS pour la production des bases de données bibligraphiques Pascal et Francis entre 1977 et 2015. L'indexation présente dans les notices est destinée à favoriser la recherche documentaire dans les bases de données bibliographiques Pascal et Francis. Elle est réalisée par des documentalistes de l'Inist qui ont une double compétence : documentaire et scientifique. Chaque documentaliste indexe ainsi les documents de son domaine de spécialité en se demandant à quelles questions le document donnera une réponse pertinente et en s'assurant que les notions appropriées figurent dans l'indexation. Il s'appuiera pour cela sur des principes garantissant une indexation de qualité (UNISIST, 1975), parmi lesquels le principe de spécificité, de conformité, d'homogénéité, d'impartialité, d'exhaustivité, tels que rappelés ci-dessous (Guinchat & Skouri, 1996) :

Spécificité : chaque document doit être indexé, si possible, au niveau le plus spécifique, en choisissant ce niveau en relation avec l'originalité du document dans le fonds documentaire et le type de questions pouvant être posées. Il peut parfois être nécessaire d'accompagner un terme spécifique de son terme générique pour le resituer dans son contexte.

Conformité : l'indexation doit se conformer au langage documentaire utilisé, tout en intégrant des mots-clés libres si le contenu du document le nécessite.

Homogénéité : le documentaliste doit s'efforcer de traiter de manière homogène le même type de documents et d'indexer les mêmes concepts toujours de la même façon.

Impartialité : l'indexation doit être le résultat d'une procédure objective qui se garde de toute évaluation personnelle.

Exhaustivité : l'indexation doit prendre en compte tous les aspects d'un document, dans la mesure où ils paraissent importants. Les concepts implicitement contenus dans le document seront également indexés afin de replacer les mots-clés dans un contexte approprié.

Pour réaliser cette tâche, les documentalistes peuvent s'appuyer sur des grilles d'indexation qui leur permettent d'identifier les notions importantes à indexer en fonction du domaine traité et de la problématique du document.

Elles servent de canevas au documentaliste qui reste seul juge pour décider de la pertinence et du poids d'une notion par rapport à la problématique de l'article. Par exemple, pour un document traitant d'un point grammatical en linguistique, le documentaliste devra rechercher le nom de la langue concernée, le nom du phénomène étudié, le domaine d'analyse, la méthodologie adoptée et le paradigme théorique dans lequel se situe l'étude.
D'autre part, afin de faciliter la tâche d'indexation face au flux important de documents à intégrer dans les bases de données, un outil de pré-indexation est proposé comme aide au documentaliste. Cet outil, développé par l'Inist, génère des mots-clés en se basant sur des règles de correspondance établies entre des mots-clés du vocabulaire contrôlé du domaine et leurs variantes à retrouver dans le texte. Par exemple, l'outil de pré-indexation proposera le mot-clé *acquisition d'une langue seconde* présent dans le référentiel ENSEIGNEMENT ET APPRENTISSAGE DES LANGUES, lorsqu'il rencontrera la forme fléchie *acquisition des langues secondes* ou la variante comportant une insertion d'un modifieur *acquisition d'une nouvelle langue seconde* dans le document à indexer. Le documentaliste intervient en validant, ou pas, les propositions d'indexation de l'outil dans le document, en les complétant le cas échéant et en mettant à jour régulièrement les référentiels contenant les règles de correspondance mot-clé/formes fléchies et variantes à rechercher.
Le documentaliste doit résoudre les problèmes liés aux notions implicites et aux ambiguïtés. Les notions implicites sont celles qui sont présentes dans le texte sans être nommées, et qui sont présentes dans le vocabulaire contrôlé. Plus le documentaliste sera expert des référentiels d'indexation, plus il sera à même de résoudre ce problème. L'ambiguïté apparaît lorsqu'un mot clé du référentiel et son occurrence dans le document différent dans leurs sens. Par exemple, un texte dans le domaine de la linguistique sur l'influence de la langue maternelle dans la perception des sons du langage comportant le mot-clé *surdité phonologique* ne devra pas être indexé avec le mot-clé *surdité* présent dans le vocabulaire contrôlé du domaine des *pathologies du langage*. Les résultats de la pré-indexation sont variables d'un domaine à l'autre. Bougouin *et al.* (2014) ont observé une échelle croissante de difficulté d'extraction de mots clés pour cinq domaines, allant de la plus facile, l'archéologie, à la plus difficile, la chimie, et des difficultés moyennes proches pour la linguistique, la psychologie et les sciences de l'information.

Les notices utilisées dans le Défi sont le résultat de ces pratiques d'indexation. Dans le cadre du projet Termith [1] dont l'objectif était l'amélioration de l'indexation automatique dans les disciplines des sciences humaines, les notices ont bénéficié d'une révision supplémentaire. Lors de celle-ci, l'accent a été mis essentiellement sur le principe de spécificité afin de s'assurer que les notices n'ont pas fait l'objet d'une indexation trop générique comme cela a pu être le cas ponctuellement, et sur le principe de conformité par rapport au langage contrôlé. En effet, le langage documentaire ayant évolué depuis la date où ont été produites les premières notices du corpus (1983), les mots-clés ont été actualisés en prenant en compte les mises à jour des vocabulaires contrôlés.

3 Indexation automatique de mots-clés

Il existe deux catégories d'indexation automatique par mots-clés : l'une libre, l'autre contrôlée. L'indexation libre consiste à extraire d'un document les unités textuelles jugées les plus importantes vis-à-vis de son contenu. L'indexation contrôlée fournit les mots-clés en se fondant sur un vocabulaire contrôlé, sans se restreindre aux unités textuelles présentes dans le document. Nous présentons en premier l'étape de sélection des mots-clés candidats, qui est une étape commune à la plupart des méthodes fournissant une indexation libre, et qui devient un objet d'étude à part entière (Wang *et al.*, 2014). Ensuite, nous présentons les méthodes d'extraction de mots-clés produisant une indexation libre, puis celles d'assignement de mots-clés produisant une indexation contrôlée.

3.1. Sélection des mots-clés candidats

La sélection des mots-clés candidats consiste à déterminer quelles sont les unités textuelles qui sont potentiellement des mots-clés, c'est-à-dire les unités textuelles qui ont des particularités similaires à celles des mots-clés définis par des humains, telles que le patron nom adjectif (par exemple, « interprétation sémantique », « concept linguistique » et « relation syntaxique »). Cette sélection réduit l'espace de recherche et permet ainsi de diminuer le temps de traitement nécessaire pour l'extraction de mots-clés et de supprimer les unités textuelles non pertinentes pouvant affecter négativement ses performances. Pour distinguer les différents candidats sélectionnés, nous

1. http://www.atilf.fr/ressources/termith/

définissons deux catégories : les candidats positifs, qui correspondent aux mots-clés assignés par des humains (mots-clés de référence), et les candidats négatifs. Parmi les candidats négatifs, nous distinguons les candidats d'importance secondaire des candidats erronés, tels que les conjonctions, les déterminants ou les unités textuelles mal segmentées (par exemple, « base du concept » issu du groupe nominal « une définition de base du concept linguistique », lui même composé des groupes nominaux « une définition de base » et « concept linguistique » dans la notice de la figure 1.1).

Il existe plusieurs méthodes de sélection de candidats, de la simple sélection de n-grammes, de *chunks* nominaux ou d'unités textuelles grammaticalement définies.

Les n-grammes sont toutes les séquences ordonnées de n mots adjacents (voir l'exemple 1). La sélection des n-grammes est très exhaustive, elle fournit un grand nombre de mots-clés candidats, ce qui maximise la quantité de candidats positifs, la quantité de candidats non importants, mais aussi la quantité de candidats erronés. Pour réduire cette dernière, il est courant de filtrer les n-grammes avec un antidictionnaire regroupant les mots ne pouvant pas être des mots-clés (conjonctions, prépositions, mots d'usage courant, etc.). Si un n-gramme contient un mot de l'antidictionnaire en début ou en fin, alors il n'est pas considéré comme un mot-clé candidat.

Malgré son aspect grossier, la sélection des n-grammes est largement utilisée en extraction de mots-clés (Witten *et al.*, 1999; Hulth, 2003; Medelyan *et al.*, 2009), pour sa simplicité de mise en œuvre.

Exemple 1 $\{1..3\}$*-grammes sélectionnés dans le phrase « L'objectif est de fournir une définition de base du concept linguistique de la cause en observant son expression. » :*

Uni-gramme	***Bi-gramme***	***Tri-gramme***
« objectif »	*« concept linguistique »*	*« définition de base »*
« définition »		*« base du concept »*
« base »		
« concept »		
« linguistique »		
« cause »		
« expression »		

Les *chunks* nominaux (*NP-chunks*) sont des syntagmes [2] non récursifs (ou minimaux) dont la tête est un nom, accompagné de ses éventuels déterminants et modifieurs usuels (voir l'exemple 2). Ils sont linguistiquement définis et leur sélection, sans considérer les déterminants qui les précèdent, est donc plus fiable que celle des n-grammes pour l'extraction de mots-clés. Hulth (2003) le montre dans ses expériences consacrées à l'apport de connaissances linguistiques pour l'extraction automatique de mots-clés. Cependant, ses propos sont nuancés par un autre de ses constats : tirer profit de la catégorie grammaticale des mots des n-grammes permet d'obtenir de meilleures performances qu'avec les *chunks* nominaux.

Exemple 2 *chunks nominaux sélectionnés dans le phrase « L'objectif est de fournir une définition de base du concept linguistique de la cause en observant son expression. » :*

Chunk nominal	***Candidat sélectionné***
« l'objectif »	*« objectif »*
« une définition »	*« définition »*
« base »	*« base »*
« concept linguistique »	*« concept linguistique »*
« la cause »	*« cause*
« expression »	*« expression »*

2. Syntagme : unité syntaxique intermédiaire entre le mot et la phrase. Aussi appelé groupe, le syntagme constitue une unité de sens dont chaque constituant conserve sa signification et sa syntaxe propre.

La sélection d'unités textuelles qui forment des séquences grammaticalement définies permet de contrôler avec précision la nature et la grammaticalité des candidats sélectionnés. Pour cela, il faut définir des patrons grammaticaux tels que `/ (N | A) +/` (voir l'exemple 3), qui représente les plus longues séquences de noms (`N`) et d'adjectifs (`A`), exprimé avec la syntaxe des expressions rationnelles.

À l'instar des *chunks* nominaux, la sélection des séquences grammaticalement définies est plus fondée linguistiquement que celle des n-grammes. Dans ses travaux, Hulth (2003) sélectionne les candidats à partir des patrons des mots-clés de référence les plus fréquents dans ses données. D'autres chercheurs, tels que Wan & Xiao (2008), se contentent des plus longues séquences de noms (noms propres inclus) et d'adjectifs.

Exemple 3 *Séquences `/ (N | A) +/` sélectionnés dans le phrase « L'objectif est de fournir une définition de base du concept linguistique de la cause en observant son expression. » :*

`/ (N | A) +/`
« objectif »
« définition »
« base »
« concept linguistique »
« cause
« expression »

3.2. Extraction de mots-clés

Les méthodes d'extraction automatique de mots-clés effectuent soit un ordonnancement par importance des mots-clés candidats vis-à-vis du contenu du document, soit une classification des mots-clés candidats entre les classes « mot-clé » et « non mot-clé ». La figure 3.2 présente la chaîne de traitements de la majorité des méthodes d'extraction de mots-clés. L'ordonnancement est principalement réalisé avec une approche non supervisée et la classification est réalisée avec une approche supervisée requérant des documents d'apprentissage manuellement indexés.

3.2..1. Méthodes non supervisées

La plupart des méthodes non supervisées d'extraction de mots-clés ordonnent les mots-clés candidats d'après leur importance vis-à-vis du contenu du document (par exemple, l'expression « concept linguistique » est importante vis-à-vis du document de la figure 1.1, page 2), puis extraient les k plus importants en tant que mots-clés. Du fait qu'elles ne requièrent pas de données d'entraînement, elles sont applicables dans toutes les situations et ont la particularité de s'abstraire du domaine des documents qu'elles traitent. Les mots-clés candidats sont analysés avec des règles simples fondées sur des traits statistiques extraits du document ou d'un corpus de référence non indexé.

TF-IDF (Salton *et al.*, 1975) et Likey (Paukkeri & Honkela, 2010) sont deux méthodes similaires qui comparent le comportement d'une unité textuelle dans le document avec son comportement dans un corpus de référence. Elles font l'hypothèse qu'une unité textuelle a une forte importance vis-à-vis du document si elle y est très fréquente et si elle l'est peu dans le corpus de référence, auquel cas elle est spécifique au document (Spärck Jones, 1972) :

$$\text{TF-IDF}(ut) = \text{TF}(ut) \times \log\left(\frac{N}{\text{DF}(ut)}\right) \quad [1]$$

$$\text{Likey}(ut) = \frac{\text{rang}_{\text{document}}(ut)}{\text{rang}_{\text{corpus}}(ut)} \quad [2]$$

Dans TF-IDF, TF (*Term Frequency*) représente le nombre d'occurrences d'une unité textuelle *ut* dans le document,

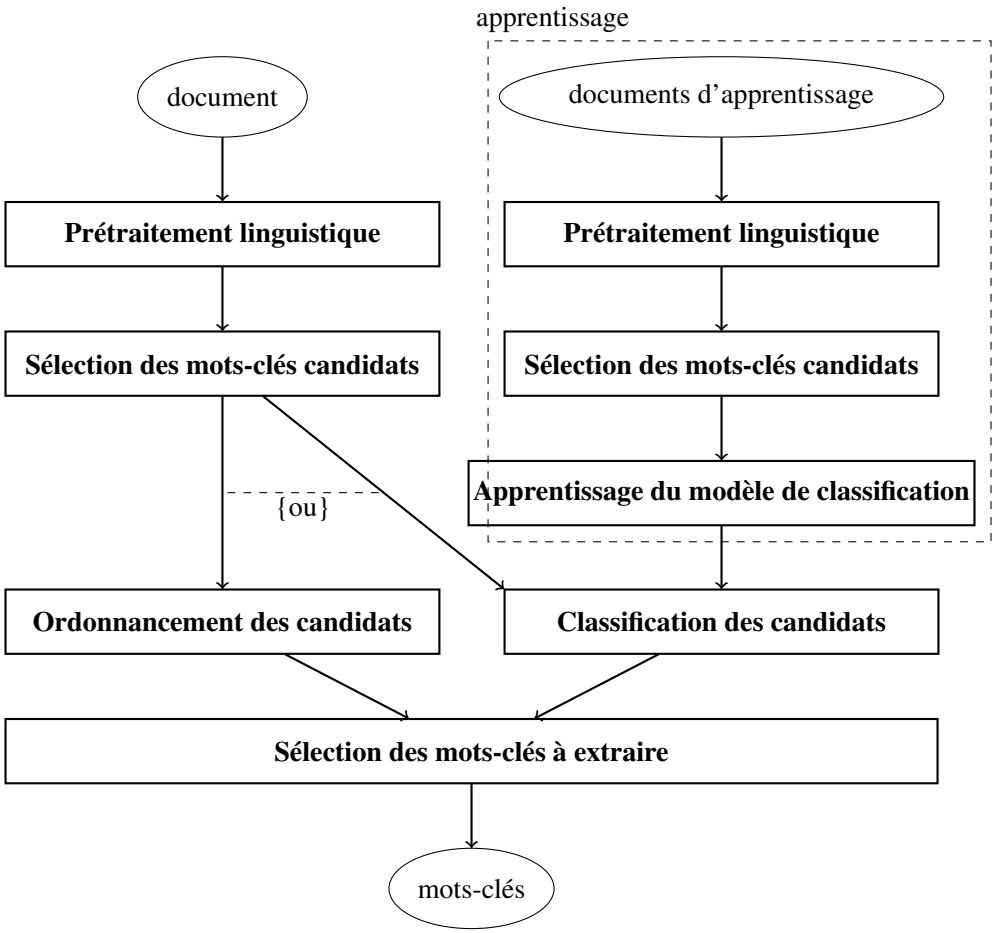

FIGURE 3.2. *Chaîne de traitements classique en extraction de mots-clés.*

DF (*Document Frequency*) représente le nombre de documents du corpus de référence dans lesquels elle occurre et N est le nombre total de documents du corpus de référence. Plus le score TF-IDF d'une unité textuelle est élevé, plus celle-ci est importante vis-à-vis du document. Dans Likey, les rangs d'une unité textuelle dans le document et dans le corpus est obtenu à partir de son nombre d'occurrences dans le document et dans le corpus, respectivement. Plus le rapport entre ces deux rangs est faible, plus l'unité textuelle évaluée est importante dans le document.

Okapi (ou BM25) (Robertson *et al.*, 1998) est une mesure alternative à TF-IDF. En RI, celle-ci est préférée à TF-IDF. Bien que l'extraction automatique de mots-clés soit une discipline entre le TAL et la RI, la méthode de pondération Okapi n'a, à notre connaissance, pas été appliquée pour l'extraction de mots-clés. Claveau (2012) décrit Okapi comme un TF-IDF prenant mieux en compte la longueur des documents. Cette dernière est utilisée pour normaliser le TF (TF_{BM25}) :

$$\text{Okapi}(ut) = \text{TF}_{\text{BM25}}(ut) \times \log\left(\frac{N - \text{DF}(ut) + 0,5}{\text{DF}(ut) + 0,5}\right) \quad [3]$$

$$\text{TF}_{\text{BM25}}(ut) = \frac{\text{TF}(ut) \times (k_1 + 1)}{\text{TF}(ut) + k_1 \times \left(1 - b + b \times \frac{\text{DL}}{\text{DL}_{\text{moyenne}}}\right)} \quad [4]$$

où k_1 est une constante fixée à 2, où b est une constante fixée à $0,75$, où DL (*Document Length*) représente la longueur du document (en nombre de mots) et où $\text{DL}_{moyenne}$ représente la longueur moyenne des documents du corpus de référence.

Outre les méthodes statistiques, des méthodes par groupement comme KeyCluster (Liu *et al.*, 2009) ou à base de graphes comme TextRank (Mihalcea & Tarau, 2004) ont été proposées mais elles peinent à égaler les performances des méthodes statistiques.

3.2..2. Méthodes supervisées

Les méthodes supervisées apprennent principalement à classer les mots-clés en tant que « mot-clé » ou « non mot-clé ». Leur apprentissage se fait à partir d'une collection d'apprentissage (ou d'entraînement) dont les documents sont manuellement indexés par des mots-clés. Les mots-clés candidats sont sélectionnés dans ces documents, ils servent d'exemples lorsqu'il font partie de l'indexation manuelle (de référence), de contre-exemples sinon et certaines de leurs caractéristiques (traits) sont analysées pour apprendre à discriminer « mots-clés » et « non mots-clés ».

Les méthodes proposées emploient des classifieurs probabilistes classiques. KEA (Witten *et al.*, 1999) est la méthode d'extraction de mots-clés la plus populaire. Elle effectue une classification naïve bayesienne pour attribuer le score de vraisemblance de chaque mot-clé candidat. Elle combine les distributions probabilistes de deux traits : la première position du candidat dans le document et son poids TF-IDF. L'intuition de Witten *et al.* (1999) est que les mots-clés ont une certaine importance vis-à-vis du document (leur poids TF-IDF) et qu'ils font leur première apparition dans des zones similaires du document.

Les champs aléatoires conditionnels (CRF) (Zhang, 2008), les arbres de décision (Turney, 2000), les séparateurs à vastes marges (SVM) (Zhang *et al.*, 2006), les perceptrons multicouche (Sarkar *et al.*, 2010) ont aussi été employés pour choisir des mots-clés parmi un ensemble de mots-clés candidats.

3.3. Assignement de mots-clés

L'assignement automatique de mots-clés assigne comme mots-clés des entrées d'un vocabulaire contrôlé indépendamment de leur présence dans celui-ci. L'assignement de mots-clés diffère donc de l'extraction de mots-clés puisque des mots clés assignés peuvent être absent du document à indexer et inversement, des mots-clés extraits peuvent ne pas appartenir à un vocabulaire contrôlé.

Medelyan & Witten (2006) ont proposé la méthode KEA++ pour effectuer de l'assignement de mots-clés. KEA++ utilise un thésaurus du domaine de spécialité utilisé d'abord pour sélectionner les mots-clés candidats, ensuite pour améliorer la classification. Medelyan & Witten (2006) réalisent l'assignement en se limitant aux mots-clés qui occurrent dans le document. Ils sélectionnent donc toutes les unités textuelles qui correspondent à une entrée du thésaurus. Ils sont ensuite classés en tant que « mot-clé » ou « non mot-clé » par un classifieur naïf bayesien. Ce classifieur est le même que celui de KEA, à l'exception d'un trait supplémentaire : le nombre de relations sémantiques qu'entretient le candidat avec les autres dans le thésaurus. De cette manière, ils déterminent l'importance du candidats dans le domaine.

4 Évaluation automatique de l'indexation de mots-clés

L'évaluation d'une méthode d'indexation automatique par mots-clés s'effectue selon un processus d'évaluation « à la Cranfield » (Voorhees, 2002). La méthode est appliquée à un ensemble de documents de test (collection de test), les mots-clés qu'elle propose pour chaque document sont comparés avec les mots-clés assignés manuellement aux documents (jugements de référence).

La mise en correspondance des mots-clés extraits/assignés aux mots-clés de référence sert à distinguer ceux qui sont corrects, les vrais positifs de ceux qui ne le sont pas, les faux positifs (cf. tableau 4.1). Toute autre unité textuelle non extraite/assignée par la méthode automatique est appelée faux négatif si elle correspond à un mot-clé de référence, et vrai négatif dans le cas contraire.

D'après le paradigme d'évaluation « à la Cranfield », un vrai positif ne peut être considéré comme tel que s'il est strictement identique à un mot-clé de référence. Cette correspondance « exacte » induit une évaluation pessimiste des méthodes d'indexation automatique par mots-clés, car les variantes linguistiques des mots-clés de référence sont jugées incorrectes sans distinction des autres faux positifs. Pour minimiser ce problème, toutes les évaluations réalisées dans la littérature tiennent compte uniquement du radical des mots des mots-clés, c'est-à-dire leur forme

		Jugement de référence	
		« mot-clé »	« non mot-clé »
Résultat	« mot-clé »	vrai positif (VP)	faux positif (FP)
	« non mot-clé »	faux negatif (FN)	vrai negatif (VN)

Tableau 4.1. *Matrice de confusion pour l'évaluation des méthodes d'indexation automatique par mots-clés.*

privée de tout suffixe (par exemple, *empir* est le radical de *empirique*). Les différences d'accords en genre et en nombre sont donc autorisées, ainsi que toute autre dérivation suffixale. Cette approche n'est pas parfaite, car elle fait parfois correspondre des mots porteurs de sens différents (par exemple, *empire* et *empirique* possèdent le même radical *empir*).

Les mesures qui ont été retenues pour la campagne sont les mesures de précision, rappel, et f1-mesure (Manning & Schütze, 1999), calculées avec une macro-moyenne. Ce sont ces mesures qui ont été utilisées pour la piste 5 de la campagne SemEval-2010 (Kim *et al.*, 2010).

La précision (P) capture la capacité d'une méthode à minimiser les erreurs. Inversement, le rappel (R) mesure la capacité de la méthode à fournir le plus possible de mots-clés corrects. Quant à la f-mesure (F), elle est un compromis entre précision et rappel, c'est-à-dire la capacité de la méthode à extraire un maximum de mots-clés corrects tout en faisant un minimum d'erreurs.

$$\mathrm{P}(d) = \frac{\#\text{NB MOTS-CLÉS EXTRAITS CORRECTS}(d)}{\#\text{NB MOTS-CLÉS EXTRAITS}(d)} \quad [5]$$

$$\mathrm{R}(d) = \frac{\#\text{NB MOTS-CLÉS EXTRAITS CORRECTS}(d)}{\#\text{NB MOTS-CLÉS DE RÉFÉRENCE}(d)} \quad [6]$$

$$\mathrm{F}(d) = 2 \times \frac{\mathrm{P}(d)\mathrm{R}(d)}{\mathrm{P}(d) + \mathrm{R}(d)} \quad [7]$$

Les résultats officiels de la campagne ont été établis sur la seule performance en f-mesure en macro-moyenne. Pour chaque méthode, les résultats de l'évaluation sont donnés par :

$$\mathrm{P} = 100 \times \frac{\sum_d \mathrm{P}(d)}{N} \quad [8]$$

$$\mathrm{R} = 100 \times \frac{\sum_d \mathrm{R}(d)}{N} \quad [9]$$

$$\mathrm{F} = 100 \times \frac{\sum_d \mathrm{F}(d)}{N} \quad [10]$$

$$[11]$$

5 Données

Les données sont composées de quatre corpus traitant chacun d'un domaine de spécialité : la linguistique, les sciences de l'information, l'archéologie et la chimie et de quatre thésaurus.

5.1. Corpus

Chaque corpus est constitué d'un ensemble de notices issues des bases de données bibliographiques Pascal et Francis de l'INIST-CNRS et qui sont fournies aux formats TEI et texte.

Chaque notice est composée de :
— un titre,
— un résumé,
— une liste de mots-clés attribuée par l'ingénieur documentaliste,
— le texte pré-traité de la notice.

La figure 5.3 donne un exemple de notice pour chaque domaine. Les textes des notices sont courts : ils ont en moyenne 156,7 mots. Quant aux mots-clés, l'indexation par des professionnels privilégie l'emploi de descripteurs appartenant à un vocabulaire contrôlé. Peu de mots-clés occurrent dans les résumés. L'exemple de notice dans le domaine de la chimie propose 25 mots-clés dont seuls deux occurrent dans le résumé. Le nombre de mots-clés varie selon les notices entre 7 mots-clés et 30. Un mot clé est généralement une unité linguistique concise, un mot simple ou une expression de deux mots qui sont tous des noms. On peut noter des spécificités par domaine : de nombreux mots-clés de l'archéologie sont des noms propres ; des formules chimiques sont employées comme mots-clés pour la chimie.

Chacun de ces corpus est divisé en deux jeux :
— Jeu d'apprentissage : ce jeu se compose de notices bibliographiques (titres et résumés), au format TEI, dans quatre domaines de spécialités explicités (linguistique, sciences de l'information, archéologie et chimie) et indexées par les indexeurs professionnels de l'INIST.
— Jeu de test (d'évaluation) : ce jeu reprend les mêmes caractéristiques que celles du jeu d'apprentissage ; la liste des mots-clés n'est n'a pas été fournie et constitue la référence pour l'évaluation.

Le corpus de linguistique est constitué de 715 notices d'articles français parus entre 2000 et 2012 dans 11 revues ; le corpus des sciences de l'information contient 706 notices d'articles français publiés entre 2001 et 2012 dans cinq revues ; le corpus d'archéologie est composé de 718 notices représentant des articles français parus entre 2001 et 2012 dans 22 revues ; le corpus de chimie est composé de 782 notices d'articles français publiés entre 1983 et 2012 dans cinq revues. Pour chaque domaine, 200 notices d'articles ont été sélectionnées au hasard pour constituer le corpus de test.

Le tableau 5.2 résume les caractéristiques du corpus d'apprentissage de chaque domaine. Pour chaque domaine de spécialité, dans la partie *Documents*, nous indiquons sous la légende *Quantité*, le nombre de notices, sous la légende *Mots moy.*, le nombre moyen de mots des notices, et sous la légende *Quantité moy.*, le nombre moyen de mots-clés associé à la notice. Toujours pour chaque domaine de spécialité, dans la partie *Mots-clés*, sous la légende *À assigner*, nous indiquons le pourcentage de mots-clés qui n'occurrent pas dans la notice, et sous la légende *Long. moy.*, la taille moyenne en nombre de mots d'un mot-clé.

Corpus	**Documents**			**Mots-clés**	
	Quantité	Mots moy.	Quantité moy.	« À assigner »	Long. moy.
Linguistique	515	160,5	8,6	61 %	1,7
Sciences de l'info.	506	105,0	7,8	68 %	1,8
Archéologie	518	221,1	16,9	37 %	1,3
Chimie	582	105,7	12,2	76 %	2,2

Tableau 5.2. *Caractéristiques des corpus d'apprentissage de DEFT.*

Nous avons aussi fourni une version analysée linguistiquement du corpus où nous avons appliqué les traitements linguistiques suivants :
— segmentation en phrases par l'outil PUNKTSENTENCETOKENIZER disponible avec la librairie Python NLTK (Bird *et al.*, 2009)
— segmentation en mots par l'outil BONSAI du BONSAI PCFG-LA PARSER 3 [3]
— étiquetage syntaxique réalisé par MElt (Denis & Sagot, 2009).

Cette mise à disposition visait à encourager les participants à utiliser ces corpus analysés plutôt que leurs propres outils afin d'évaluer plutôt les algorithmes d'indexation que les traitements du TALN.

3. https://raweb.inria.fr/rapportsactivite/RA2011/alpage/uid47.html

La cause linguistique *Linguistique*

L'objectif est de fournir une définition de base du concept linguistique de la cause en observant son expression. Dans un premier temps, l'A. se demande si un tel concept existe en langue. Puis il part des formes de son expression principale et directe (les verbes et les conjonctions de cause) pour caractériser linguistiquement ce qui fonde une telle notion.

Mots-clés : français ; interprétation sémantique ; conjonction ; expression linguistique ; concept linguistique ; relation syntaxique ; cause.

Congrès de l'ABF : les publics des bibliothèques *Sciences de l'info.*

Le cinquante-troisième congrès annuel de l'Association des bibliothécaires de France (ABF) s'est déroulé à Nantes du 8 au 10 juin 2007. Centré sur le thème des publics, il a notamment permis de méditer les résultats de diverses enquêtes auprès des usagers, d'examiner de nouvelles formes de partenariats et d'innovations technologiques permettant aux bibliothèques de conquérir de nouveaux publics, et montré des exemples convaincants d'ouverture et d'"hybridation", conditions du développement et de la fidélisation de ces publics.

Mots-clés : rôle professionnel ; évolution ; bibliothèque ; politique bibliothèque ; étude utilisateur ; besoin de l'utilisateur ; partenariat ; web 2.0 ; centre culturel.

Étude préliminaire de la céramique non tournée micacée du bas Languedoc occidental : typologie, chronologie et aire de diffusion *Archéologie*

L'étude présente une variété de céramique non tournée dont la typologie et l'analyse des décors permettent de l'identifier facilement. La nature de l'argile enrichie de mica donne un aspect pailleté à la pâte sur laquelle le décor effectué selon la méthode du brunissoir apparaît en traits brillant sur fond mat. Cette première approche se fonde sur deux séries issues de fouilles anciennes menées sur les oppidums du Cayla à Mailhac (Aude) et de Mourrel-Ferrat à Olonzac (Hérault). La carte de répartition fait état d'échanges ou de commerce à l'échelon macrorégional rarement mis en évidence pour de la céramique non tournée. S'il est difficile de statuer sur l'origine des décors, il semble que la production s'insère dans une ambiance celtisante. La chronologie de cette production se situe dans le deuxième âge du Fer. La fourchette proposée entre la fin du IV^{e} et la fin du II^{e} s. av. J.-C. reste encore à préciser.

Mots-clés : distribution ; mourrel-ferrat ; olonzac ; le cayla ; mailhac ; micassé ; céramique non-tournée ; celtes ; production ; echange ; commerce ; cartographie ; habitat ; oppidum ; site fortifié ; fouille ancienne ; identification ; décor ; analyse ; répartition ; diffusion ; chronologie ; typologie ; céramique ; étude du matériel ; hérault ; aude ; france ; europe ; la tène ; age du fer.

Réaction entre solvant et espèces intermédiaires apparues lors de l'électroréduction-acylation de la fluorénone et de la fluorénone-anil dans l'acétonitrile *Chimie*

Étude du comportement des différents acylates de fluorénols-9 vis-à-vis des anions CH_2CN (électrogénérés par réduction de l'azobenzène en son dianion dans l'acétonitrile). Réduction de la fluorénone dans l'acétonitrile en présence de chlorures d'acides ou d'anhydrides

Mots-clés : réduction chimique ; acylation ; réaction électrochimique ; acétonitrile ; composé aromatique ; composé tricyclique ; cétone ; cétimine ; effet solvant ; effet milieu ; radical libre organique anionique ; mécanisme réaction ; nitrile ; hydroxynitrile ; composé saturé ; composé aliphatique ; anhydride organique ; fluorénone ; fluorénone,phénylimine ; fluorénol-9,acylate ; fluorènepropiononitrile-9(hydroxy-9) ; bifluorényle-9,9pdiol-9,9p ; fluorèneδ9 :α-acétonitrile ; butyrique acide(chloro-4) chlorure.

FIGURE 5.3. *Exemple de notices Termith pour chaque domaine. Les mots-clés soulignés occurrent dans la notice.*

5.2. Référentiels

Les référentiels correspondent aux vocabulaires contrôlés utilisés pour l'indexation des bases de données bibliographiques de l'INIST-CNRS.

Le vocabulaire contrôlé est une liste de mots-clés possibles dans un domaine de spécialité. Cette liste est plus ou moins structurée en fonction des domaines. Les mots-clés sont mis en relations s'ils sont associés à un même concept (par exemple, « nom composé » et « substantif composé » en linguistique) ou si l'un est l'hyperonyme de l'autre, c'est-à-dire plus générique (par exemple « allemand » par rapport à « haut-allemand » et « bas-allemand »).

En définissant le langage documentaire à utiliser pour indexer les documents du même domaine, le vocabulaire contrôlé contribue à la conformité et à l'homogénéité de l'indexation. Il n'assure cependant pas l'exhaustivité

Domaine	Total entrées	Composition	
		Vocabulaire contrôlé	Volume entrées
Linguistique	13 968	ML (sciences du langage)	6 079
		MC (sciences de l'éducation)	2 681
		MS (sociologie)	5 208
Sciences de l'info.	92 472	MX (Sciences exactes, sciences de l'ingénieur et technologies)	92 472
Archéologie	4 905	MA (art et archéologie)	1 849
		MH (préhistoire et protohistoire)	3 056
Chimie	122 359	MX (Sciences exactes, sciences de l'ingénieur et technologies)	92 472
		M3 (Physique)	29 887

Tableau 5.3. *Caractéristiques des thésaurus.*

et doit être mis à jour régulièrement, soit par une veille terminologique, soit au fur et à mesure des indexations manuelles, pour intégrer les nouveaux concepts.

Pour le défi, certains domaines ont fait l'objet d'un regroupement de vocabulaires afin de se rapprocher de la couverture du corpus de notices, par exemple, en archéologie, regroupement de deux vocabulaires (MA – MH), en linguistique, regroupement de trois vocabulaires (ML – MC – MS) et en chimie, regroupement de deux vocabulaires (MX – M3). D'autres vocabulaires sont quant à eux inclus dans un seul vocabulaire très multidisciplinaire (MX), c'est le cas pour les sciences de l'information et la chimie. Le détail des regroupements de vocabulaires est donné dans le tableau 5.3.

Les vocabulaires contrôlés ou référentiels, associés à chaque domaine de spécialité ont été fournis au format SKOS (Simple Knowledge Organization System). La figure 5.4 montre un extrait de thésaurus dans ce format. Les entrées du thésaurus sont les balises `Concept`. Chaque concept possède un identifiant de concept (l'attribut RDF :ABOUT), une sous-balise PREFLABEL donnant l'étiquette principale du concept (le mot préférentiel), et éventuellement une ou plusieurs sous-balises ALTLABEL donnant les étiquettes alternatives du concept (les synonymes ou les anciens préférentiels). Comme stipulé dans la spécification SKOS, les concepts peuvent également posséder des sous-balises indiquant des relations sémantiques entre eux. Par exemple, la balise BROADER renvoie vers un concept générique. La balise RELATED renvoie vers un concept associé. La documentation des balises sémantiques du format SKOS est donnée par la section 8 des spécifications SKOS [4].

6 Campagne

Un appel à participation a été lancé le 15 janvier 2016 sur les principales listes du traitement automatique des langues. Huit équipes se sont inscrites et cinq équipes ont participé aux tests. Ces équipes sont les suivantes :

EBSI *École de Bibliothéconomie et des Sciences de l'Information, Université de Montréal : Dominic Forest, Jean-François Chartier et Olivier Lacombe*

EXENSA *SAS eXenSa* [5] *: Morgane Marchand*

LIMSI *Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur : Thierry Hamon*

LINA *Laboratoire d'Informatique de Nantes Atlantique, Université de Nantes : Adrien, Bougouin, Florian Boudin et Béatrice Daille*

LIPN *Laboratoire d'Informatique de Paris Nord, Université Paris 13 : Haïfa Zargayouna et Davide Buscaldi*

Les corpus d'apprentissage ont été diffusés le 2 mars 2016 aux participants, avec le script d'évaluation que nous

4. https://www.w3.org/TR/2009/REC-skos-reference-20090818/#semantic-relations
5. http://www.exensa.com/

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:dc="http://purl.org/dc/elements/1.1"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
         xmlns:owl="http://www.w3.org/2002/07/owl#"
         xmlns:dct="http://purl.org/dc/terms/"
         xmlns="http://www.w3.org/2004/02/skos/core#">
  <owl:Ontology>
     <dct:title>
        Controlled vocabulary extracted from
        INIST-CNRS database
     </dct:title>
     <dct:rightsHolder>
        INIST-CNRS (Institut de l'Information Scientifique et Technique -
        CentreNational de la Recherche scientifique)
     </dct:rightsHolder>
     <dct:dateCopyrighted>February 14, 2016</dct:dateCopyrighted>
     <dct:license rdf:about="http://creativecommons.org/licenses/by/4.0/">
       <p>
            The Creative Commons Attribution 4.0 International
            License applies to this document.
       </p>
       <p>
            Any re-use of this resource should attribute its
            content to <q>INIST-CNRS</q>
       </p>
     </dct:license>
   </owl:Ontology>
  <Concept rdf:about="http://www.inist.fr/basevoc/archeologie#ma_97563">
    <prefLabel xml:lang="fr">Abandon de site</prefLabel>
  </Concept>
  <Concept rdf:about="http://www.inist.fr/basevoc/archeologie#ma_97565">
    <prefLabel xml:lang="fr">Abeille</prefLabel>
  </Concept>
  <Concept rdf:about="http://www.inist.fr/basevoc/archeologie#ma_97566">
    <prefLabel xml:lang="fr">Abri</prefLabel>
  </Concept>
  <Concept rdf:about="http://www.inist.fr/basevoc/archeologie#ma_97567">
    <prefLabel xml:lang="fr">Académie</prefLabel>
  </Concept>
  <Concept rdf:about="http://www.inist.fr/basevoc/archeologie#ma_97569">
    <prefLabel xml:lang="fr">Acier</prefLabel>
  </Concept>
  <Concept rdf:about="http://www.inist.fr/basevoc/archeologie#ma_97570">
    <prefLabel xml:lang="fr">Objet en acier</prefLabel>
    <altLabel xml:lang="fr">Acier objet</altLabel>
  </Concept>
  ...
</rdf:RDF>

...
```

FIGURE 5.4. *Extrait de thésaurus au format SKOS.*

	Moy(Préc.)	Moy(Rap.)	Moy(f-score)
DEFT	24, 92	30, 40	**25, 03**
Tf-IDF	17,7	16,375	15,78
KEA++	14	11,975	12,425

Tableau 7.4. *Précision, rappel et f-score moyens des meilleures méthodes de chaque équipe.*

Rang	Équipe candidate	Points
1er	**eXenSa**	**18**
2ième	EBSI	16
3ième	LINA	12
4ième	LIMSI	7
4ième	LIPN	7

Tableau 7.5. *Classement général de DEFT2016.*

avons utilisé pour calculer les scores finaux [6]. Les participants ont bénéficié de six semaines pour élaborer sur les jeux d'apprentissage un maximum de trois méthodes d'extraction `m1`, `m2` et `m3`. Pour la phase de test, les équipes participantes ont chacune disposé d'une plage de trois jours choisie selon leurs disponibilités dans la semaine du 11 au 17 avril 2016. Les jeux de test leur ont été fournis individuellement par le comité d'organisation au début de cette période et les participants ont retourné dans un délai de 72 h les mots-clés extraits par chacune de leurs trois méthodes et pour chacun des quatre corpus. Ce sont donc douze fichiers de résultats que chaque participant était autorisé à produire.

Les méthodes proposées par les participants sont toutes originales et adoptent des approches différentes. EBSI et EXensa proposent des approches entièrement numériques employant une modélisation vectorielle du document. LINA propose une méthode d'ordonnancement à partir d'un graphe. LIMSI et LIPN proposent tous les deux des méthodes symboliques relevant du traitement automatique des langues. LIMSI effectue une extraction de mots-clés à l'aide d'un programme d'extraction terminologique qui regroupe les mots-clés en fonction de relations qu'ils entretiennent entre eux, puis les sélectionne à partir de leur position dans le texte et du vocabulaire qui les compose. LIPN est la seule méthode qui assigne automatiquement les mots-clés en exploitant les référentiels fournis. Aucune des méthodes à l'exception du LINA n'a utilisée la version du corpus analysée linguistiquement. LIPN et LIMSI ont employé leur propre chaîne de traitements. EBSI et EXensa ont appliqué des traitements linguistiques se résumant à une normalisation des chaînes de caractères.

7 Résultats

Pour chaque corpus, seule la meilleure méthode en f-score de chaque équipe a été retenue (*cf.* section 7.2.). Le tableau 7.4 illustre la difficulté de la tâche en produisant la moyenne des f-score des meilleures méthodes de chaque équipe. **Le f-score général moyen est de 25, 03** %. Comparé aux deux méthodes de référence présentées en Section 3, TF-IDF pour l'extraction de mots-clés et KEA++ pour l'assignement de mots-clés, les méthodes proposées sont bien plus performantes.

7.1. Classement général

L'équipe candidate qui arrive en tête du concours DEFT2016 est l'équipe eXenSa.

6. Bien que ce script ait fait l'objet entre-temps d'une légère modification pour corriger un problème avec le corpus « linguistique »

Rang	Méthode	Moy(Préc.)	Moy(Rap.)	Moy(F-mesure)
1ier	exensa-m1	28, 24	34, 37	**29, 30**
2ième	ebsi-m2	27, 44	33, 05	**29, 13**
3ième	ebsi-m1	27, 73	32, 24	**28, 88**
4ième	ebsi-m3	25, 78	30, 85	**27, 28**
5ième	lina-m3	30, 00	24, 67	**26, 01**
6ième	lina-m1	28, 39	23, 53	**24, 71**
7ième	limsi-m2	25, 75	20, 23	**21, 65**
8ième	limsi-m1	24, 31	21, 88	**21, 42**
9ième	limsi-m3	25, 24	19, 79	**21, 20**
10ième	lipn-m3	13, 28	39, 66	**19, 04**
11ième	lina-m2	22, 21	17, 79	**18, 91**
12ième	lipn-m1	16, 67	21, 59	**17, 12**
13ième	lipn-m2	14, 12	24, 03	**17, 11**

t

Tableau 7.6. *Classement exhaustif de toutes méthodes proposées par tous les participants.*

#	Candidat	Préc.	Rap.	F-mesure	Points
1.	ebsi-m2	30, 26	34, 16	**31, 75**	5
2,	exensa-m1	23, 28	32, 73	**26, 30**	4
3,	lina-m3	23, 16	25, 85	**24, 19**	3
4,	lipn-m2	13, 98	30, 81	**19, 07**	2
5,	limsi-m2	15, 67	16, 10	**15, 63**	1

Tableau 7.7. *Linguistique.*

7.1..1. Classement général des équipes candidates

Le classement général des équipes est obtenu en retenant pour chaque corpus et pour chaque équipe candidate la meilleure méthode en f-score. Ces classements sont publiés en section 7.2.. Pour chaque corpus, 5 points sont attribués à l'équipe qui arrive en tête, puis 4 à la deuxième, et ainsi de suite. Le total des points donne le classement général est donné par le tableau 7.5.

7.1..2. Classement général des méthodes

Le classement général des méthodes (*cf.* tableau 7.6) donne le positionnement global de chaque méthode candidate. Le score de chaque méthode est obtenu en effectuant une moyenne des quatre valeurs de f-score obtenues pour chacun des quatre corpus. Nous pouvons aussi observer la faible performance des méthodes d'extraction de mots-clés avec une f-mesure moyenne de 25 %. Ceci peut s'expliquer par l'évaluation automatique stricte qui n'accepte pas les correspondances partielles comme pour les mots-clés *articles* et *articles de recherche* qui pour une notice peuvent être employé indifféremment.

7.2. Classement f-score par corpus

Les classements spécifiques à chacun des quatre corpus : *Linguistique* (tableau 7.7), *Sciences-info* (tableau 7.8), *Archéologie* (tableau 7.9) et *Chimie* (tableau 7.10) sont produits en ne retenant que la meilleure méthode en f-score de chaque équipe candidate. Les scores obtenus par les méthodes montrent des écarts élevés entre les domaines : l'archéologie apparaît comme le domaine le plus facile à indexer, la chimie le plus difficile, les sciences de l'information et la linguistique entre ces deux bornes. Ce constat avait déjà été fait par Bougouin *et al.* (2014), il est confirmé par l'ensemble des méthodes.

t !

#	Candidat	Préc.	Rap.	F-mesure	Points
1.	ebsi-m1	31, 03	28, 23	**28, 98**	5
2,	exensa-m1	21, 26	30, 32	**23, 86**	4
3,	lina-m3	21, 93	21, 83	**21, 45**	3
4,	lipn-m2	11, 72	23, 54	**15, 34**	2
5,	limsi-m2	13, 83	12, 01	**12, 49**	1

Tableau 7.8. *Sciences-info.*

#	Candidat	Préc.	Rap.	F-mesure	Points
1.	exensa-m1	43, 48	52, 71	**45, 59**	5
2,	limsi-m3	55, 26	38, 03	**43, 26**	4
3,	lina-m3	53, 77	33, 46	**40, 11**	3
4,	ebsi-m2	30, 77	43, 24	**34, 96**	2
5,	lipn-m1	33, 93	31, 25	**30, 75**	1

Tableau 7.9. *Archéologie.*

#	Candidat	Préc.	Rap.	F-mesure	Points
1.	exensa-m1	24, 92	21, 73	**21, 46**	5
2,	ebsi-m2	19, 67	25, 07	**21, 07**	4
3,	lina-m3	21, 15	17, 54	**18, 28**	3
4,	lipn-m3	10, 88	30, 25	**15, 31**	2
5,	limsi-m2	18, 19	14, 90	**15, 29**	1

Tableau 7.10. *Chimie.*

8 Conclusion

L'indexation d'articles scientifiques est une tâche ancienne au carrefour de la recherche d'information et du traitement automatique des langues. L'objectif de ce défi était de simuler l'indexation réalisée par des indexeurs professionnels qui s'appuient sur des thésaurus du domaine de spécialité et sur la notice de l'article. Quatre domaines de spécialité ont été expérimentés pour le français : linguistique, sciences de l'information, archéologie et chimie. Le défi a été relevé par cinq participants qui ont tous proposés des méthodes différentes et originales : methode d'extraction de mots-clés, méthode d'assignement de mots-clés et méthodes numériques ou à base de graphe s'appuyant sur une modélisation des données d'apprentissage. Malgré son ancienneté, l'indexation d'articles scientifiques reste une tâche difficile. Même si les méthodes proposées fournissent des résultats bien au-dessus des méthodes état de l'art pour l'extraction ou l'indexation de mots-clés, la f-mesure moyenne des meilleures méthodes des participants reste à 25,3 %. De plus, il existe des écarts élevés entre les domaines : l'archéologie apparaît comme le domaine le plus facile à indexer, la chimie le plus difficile. Ces résultats sont néanmoins à prendre avec circonspection, l'évaluation automatique s'effectuant sur une identité stricte calculée sur les radicaux calculés automatiquement des mots-clés candidats et des mots-clés de référence. L'amélioration de la tâche d'indexation devra sans doute passer par l'exploitation du texte intégral, ce qui pourra constituer une nouvelle édition du défi DEFT d'indexation d'articles scientifiques ou par la mise en œuvre de méthodes d'évaluation plus sophistiquées qui s'appuieraient sur les lemmes ou recourreraient à l'emploi d'une distance sémantique.

9 Remerciement

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence (ANR-12-CORD-0029).

Références

BIRD S., KLEIN E. & LOPER E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

BOUGOUIN A., BOUDIN F. & DAILLE B. (2014). Influence des domaines de spécialité dans l'extraction de termes-clés. In *Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles*, p. 13–24, Marseille, France : Association pour le Traitement Automatique des Langues.

CLAVEAU V. (2012). Vectorisation, Okapi et calcul de similarité pour le TAL : pour oublier enfin le TF-IDF (Vectorization, Okapi and Computing Similarity for NLP : Say Goodbye to TF-IDF) [in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, Volume 2 : TALN*, p. 85–98, Grenoble, France : ATALA/AFCP.

DENIS P. & SAGOT B. (2009). Coupling an Annotated Corpus and a Morphosyntactic Lexicon for State-of-the-Art POS Tagging with Less Human Effort. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC)*, p. 110–119, Hong Kong : City University of Hong Kong.

GUINCHAT C. & SKOURI Y. (1996). *Guide pratique des techniques documentaires*. Vanves : EDICEF.

HULTH A. (2003). Improved Automatic Keyword Extraction Given More Linguistic Knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 216–223, Stroudsburg, PA, USA : Association for Computational Linguistics.

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 (SemEval)*, p. 21–26, Stroudsburg, PA, USA : Association for Computational Linguistics.

LIU Z., LI P., ZHENG Y. & SUN M. (2009). Clustering to Find Exemplar Terms for Keyphrase Extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 1 (EMNLP)*, p. 257–266, Stroudsburg, PA, USA : Association for Computational Linguistics.

MANNING C. D. & SCHÜTZE H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA : MIT Press.

MEDELYAN O., FRANK E. & WITTEN I. H. (2009). Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, p. 1318–1327, Singapore : Association for Computational Linguistics.

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 : ACM.

MIHALCEA R. & TARAU P. (2004). TextRank : Bringing Order Into Texts. In DEKANG LIN & DEKAI WU, Eds., *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 404–411, Barcelona, Spain : Association for Computational Linguistics.

PAROUBEK P., ZWEIGENBAUM P., FOREST D. & GROUIN C. (2012). Indexation libre et contrôlée d'articles scientifiques. Présentation et résultats du défi fouille de textes DEFT2012 (Controlled and Free Indexing of Scientific Papers. Presentation and Results of the DEFT2012 Text-Mining Challenge) [in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, Workshop DEFT 2012 : DÉfi Fouille de Textes (DEFT 2012 Workshop : Text Mining Challenge)*, p. 1–13, Grenoble, France : ATALA/AFCP.

PAUKKERI M.-S. & HONKELA T. (2010). Likey : Unsupervised Language-Independent Keyphrase Extraction. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval)*, p. 162–165, Stroudsburg, PA, USA : Association for Computational Linguistics.

ROBERTSON S. E., WALKER STEVE & HANCOCK-BEAULIEU MICHELINE (1998). Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive Track. In *Proceedings of the Text REtrieval Conference (TREC)*, p. 199–210.

SALTON G., WONG A. & YANG C. (1975). A Vector Space Model for Automatic Indexing. *Communication ACM*, 18(11), 613–620.

SARKAR K., NASIPURI M. & GHOSE S. (2010). A New Approach to Keyphrase Extraction Using Neural Networks. *International Journal of Computer Science Issues Publicity Board 2010*.

SPÄRCK JONES K. (1972). A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 28(1), 11–21.

TURNEY P. D. (2000). Learning Algorithms for Keyphrase Extraction. *Information Retrieval*, 2(4), 303–336.

UNISIST (1975). *Principes d'indexation*. Unesco, Paris. (SC/75/WS/58).

VOORHEES E. M. (2002). The Philosophy of Information Retrieval Evaluation. In *Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, p. 355–370, London, UK : Springer-Verlag.

WAN X. & XIAO J. (2008). Single Document Keyphrase Extraction Using Neighborhood Knowledge. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, p. 855–860 : AAAI Press.

WANG R., LIU W. & MCDONALD C. (2014). How Preprocessing Affects Unsupervised Keyphrase Extraction. In A. GELBUKH, Ed., *Computational Linguistics and Intelligent Text Processing*, volume 8403 of *Lecture Notes in Computer Science*, p. 163–176. Springer Berlin Heidelberg.

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.

ZHANG C. (2008). Automatic keyword extraction from documents using conditional random fields. *Journal of Computational Information Systems*, 4(3), 1169–1180.

ZHANG K., XU H., TANG J. & LI J. (2006). Keyword Extraction Using Support Vector Machine. In *Proceedings of the 7th International Conference on Advances in Web-Age Information Management*, p. 85–96, Berlin, Heidelberg : Springer-Verlag.

