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

Béatrice Daille* Sabine Barreaux† Florian Boudin* Adrien Bougouin*
Damien Cram* Amir Hazem*
*LINA – UMR CNRS 6241, 2 rue de la Houssinière, 44322 Nantes Cedex 3, France
† INIST CNRS, 2, allée du Parc de Brabois, 54519 Vandœuvre-lès-Nancy, France
<prenom.nom>@univ-nantes.fr, <prenom.nom>@inist.fr

RÉSUMÉ

Nous présentons la campagne 2016 du défi fouille de textes (DEFT), qui pour sa douzième édition a proposé aux participants de travailler sur la problématique de l'indexation de documents scientifiques. La tâche a consisté à indexer à l'aide de mots-clés des notices bibliographiques, en français, dans quatre domaines de spécialité (linguistique, sciences de l'information, archéologie et chimie) et dont l'indexation de référence a été réalisée par des indexeurs professionnels. Les résultats ont été évalués avec les mesures de précision, rappel, et f1-mesure, calculés avec une macro-moyenne.

ABSTRACT

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

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 on four domains: linguistics, information Science, archaeology and chemisty. The results have been evaluated in terms of precision, recall and f-measure computed after stemming upon the reference 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), nous proposons de travailler sur l'indexation de documents scientifiques par l'intermédiaire de mots-clés. Alors que l'édition 2012 visait l'identification des mots-clés d'auteurs, nous avons proposé de travailler sur l'identification des mots-clés proposés par des indexeurs professionnels (ingénieurs documentalistes).

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 des mots-clés, nous parlons d'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, nous parlons d'indexation libre.

Les méthodes devront identifier les concepts importants permettant d'indexer les documents. Comme l'indexation proposée par les indexeurs professionnels, les méthodes pourront proposer une indexation contrôlée, libre ou mixte.

2 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.

2.1 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 1 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 1 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 Á 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

Table 1: Caractéristiques des corpus d'apprentissage de DEFT

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

Table 2: Caractéristiques des thésaurus

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 1: Exemple de notices Termith pour chaque domaine. Les mots-clés soulignés occurrent dans la notice.

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[1]
- 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.

2.2 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 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 2.

Les vocabulaires contrôlés ou référentiels, associés à chaque domaine de spécialité ont été fournis au format SKOS (Simple Knowledge Organization System). La figure 2 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 terme 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[2].

[1] `https://raweb.inria.fr/rapportsactivite/RA2011/alpage/uid47.html`
[2] `https://www.w3.org/TR/2009/REC-skos-reference-20090818/#semantic-relations`

```
<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 2: Extrait de thésaurus au format SKOS

3 Tâche proposée

La tâche consiste à fournir pour une notice bibliographique (titre + résumé) les mots-clés la caractérisant au mieux. Cette tâche simule l'indexation réalisée par un professionnel, qui s'appuie sur des référentiels (des thesaurus), et éventuellement complète la liste issue des référentiels par des mots-clés apparaissant ou non dans la notice. Les données porteront sur quatre domaines de spécialité (linguistique, sciences de l'information, archéologie et chimie). L'indexation de référence a été revue dans le cadre du projet TermiTH[3].

4 Évaluation

Les mesures qui ont été retenues pour l'évaluation 2016 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.

$$\mathrm{P}(d) = \frac{\#\text{NB MOTS-CLÉS EXTRAITS CORRECTS}(d)}{\#\text{NB MOTS-CLÉS EXTRAITS}(d)} \tag{1}$$

$$\mathrm{R}(d) = \frac{\#\text{NB MOTS-CLÉS EXTRAITS CORRECTS}(d)}{\#\text{NB MOTS-CLÉS DE RÉFÉRENCE}(d)} \tag{2}$$

$$\mathrm{F}(d) = 2 \times \frac{\mathrm{P}(d)\mathrm{R}(d)}{\mathrm{P}(d) + \mathrm{R}(d)} \tag{3}$$

Pour comparer les mots-clés fournis par les participants à la référence, nous avons utilisé l'égalité stricte sur les mots-clés. Afin de ne pas biaiser l'évaluation par rapport à une ontologie particulière, nous avons décidé de ne pas recourir à l'emploi d'une distance sémantique qui permettrait par exemple de s'apercevoir que *recherche d'information* est plus proche de *fouille de données* que d'*algorithmique*, ni de prendre en compte les recouvrements partiels de mots-clés comme ayant une certaine validité pour éviter de récompenser un système qui retournerait *fouilles archéologiques* alors que la bonne réponse est *fouille de données*. Bien entendu, ce choix a pour résultat que, par exemple, l'identification d'un hyponyme d'un mot-clé au lieu du mot-clé sera considérée comme aussi fausse que l'identification de n'importe quel autre mot. En revanche, nous acceptons les variantes flexionnelles.

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 :

[3] http://www.atilf.fr/ressources/termith/

$$\mathrm{P} = 100 \times \frac{\sum_d \mathrm{P}(d)}{N} \tag{4}$$

$$\mathrm{R} = 100 \times \frac{\sum_d \mathrm{R}(d)}{N} \tag{5}$$

$$\mathrm{F} = 100 \times \frac{\sum_d \mathrm{F}(d)}{N} \tag{6}$$

$$\tag{7}$$

5 Résultats

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 :

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*

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

EXENSA *SAS eXenSa*[4] *: Morgane Marchand*

Les corpus d'apprentissage ont été diffusés le 2 mars 2016 aux participants, avec le script d'évaluation que nous avons utilisé pour calculer les scores finaux[5]. 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 72h 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. Pour chaque corpus, seule la meilleure méthode en f-score de chaque équipe a été retenue (*cf.* section 5.2). Le tableau 3 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 $25,03\,\%$.**

[4] `http://www.exensa.com/`

[5] Bien que ce script ait fait l'objet entre-temps d'une légère modification pour corriger un problème avec le corpus " linguistique"

Moy(Préc.)	Moy(Rap.)	**Moy(f-score)**
24.92	30.40	**25.03**

Table 3: 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

Table 4: Classement général de DEFT2016

5.1 Classement général

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

5.1.1 Classement général des équipes candidates

Le classement général des équipes est obtenu en ne retenant pour chaque corpus et pour chaque équipe candidate que la meilleure méthode en f-score. Ces classements sont publiés en section 5.2. 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 4.

5.1.2 Classement général des méthodes

Le classement général des méthodes (*cf.* tableau 5) 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 (p. ex. *articles* et *articles de recherche* qui en contexte réfèrent au même concept.

5.2 Classement f-score par corpus

Les classements spécifiques à chacun des quatre corpus : *Linguistique* (tableau 6), *Sciences-info* (tableau 7), *Archéologie* (tableau 8) et *Chimie* (tableau 9) 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.

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**

Table 5: 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

Table 6: Linguistique

#	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

Table 7: 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

Table 8: 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

Table 9: Chimie

6 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 : linguistique, sciences de l'information, archéologie et chimie. Malgré son ancienneté, l'indexation d'articles scientifiques reste une tâche difficile, la f-mesure moyenne étant de 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. L'amélioration de la tâche d'indexation devra sans doute passer par l'exploitation du texte plein, ce qui pourra constituer une nouvelle édition du défi DEFT d'indexation d'articles scientifiques.

Remerciements

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.

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.

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.

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

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.