AFCP

ATALA

JEP-TALN-RECITAL 2016

Journées d'Études sur la Parole
Traitement Automatique des Langues Naturelles
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues

PARIS Inalco du 4 au 8 juillet 2016

Organisé par les laboratoires franciliens

https://jep-taln2016.limsi.fr

Conférenciers invités:

Christian Chiarcos (Goethe-Universität, Frankfurt.)
Mark Liberman (University of Pennsylvania, Philadelphia)

Coordinateurs comités d'organisation

Nicolas Audibert et Sophie Rosset (JEP)
Laurence Danlos & Thierry Hamon (TALN)
Damien Nouvel & Ilaine Wang (RECITAL)
Philippe Boula de Mareuil, Sarra El Ayari & Cyril Grouin (Ateliers)

©2016 Association Francophone pour la Communication Parlée (AFCP) et
Association pour le Traitement Automatique des Langues (ATALA)

Table des matières

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 1

Alignement de deux espaces sémantiques à des fins d'indexation automatique
Jean-François Chartier, Dominic Forest, Olivier Lacombe 13

Indexation automatique de notices bibliographiques à l'aide d'approches d'acquisition terminologique
Thierry Hamon 20

LIPN@DEFT2016 : Annotation de documents en utilisant l'Information Mutuelle
Davide Buscaldi, Haïfa Zargayouna 27

Représentation vectorielle de mots pour l'indexation de notices bibliographiques
Morgane Marchand, Geoffroy Fouquier, Guillaume Pitel 34

TopicRank en domaines de spécialité : participation du LINA à DEFT 2016
Adrien Bougouin, Florian Boudin, Béatrice Daille 41

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.

Alignement de deux espaces sémantiques à des fins d'indexation automatique

Jean-François Chartier, Dominic Forest et Olivier Lacombe
Université de Montréal, École de bibliothéconomie et des sciences de l'information,
C.P. 6128, succursale Centre-ville, Montréal (Québec), H3C 3J7, Canada
Chartier.jf@gmail.com, dominic.forest@umontreal.ca,
lacombe.olivier.ol@gmail.com

RESUME

Cet article présente la méthode et les résultats de l'équipe de l'Université de Montréal à la 12e édition du Défi Fouille de Textes. La méthode développée repose sur une procédure d'apprentissage automatique supervisée. Elle est basée sur un espace sémantique des mots-clés d'indexation (ESMC) induit à partir de la base d'apprentissage et d'un espace sémantique de documents construit à partir de la base de test (ESD). La prédiction des mots-clés d'indexation pour un document de la base de test est réalisée en calculant la proximité entre les documents de l'ESD et les mots-clés de l'ESMC. Les k mots-clés les plus proches d'un document sont considérés être les mots-clés les plus pertinents pour son indexation.

ABSTRACT

This article presents the method and the results of the Université de Montréal team in the 12th edition of the Text Mining Challenge. The method developed is a supervised machine learning procedure. It is based on a keyword semantic space (ESMC) induced from the training set and a document semantic space (ESD) built from the test set (ESD). The keywords prediction for the test set is performed by calculating the proximity between the documents of the ESD and the keywords of the ESMC. The k nearest keywords of a document are considered to be the most relevant keywords for the document indexation.

MOTS-CLES : Indexation automatique ; Espace sémantique ; Corrélation ; Assignation de mots-clés ; apprentissage automatique

KEYWORDS: Automatic indexation ; Semantic Vector Space ; Correlation ; Keywords assignment ; Machine Learning

1 Introduction

La masse documentaire disponible en format numérique augmente rapidement. Cette réalité rend l'indexation manuelle des documents de plus en plus difficile à maintenir. Pour gérer cette masse documentaire des méthodes d'indexation automatique ont été proposées. Deux types de méthodes existent pour attribuer automatiquement des mots-clés d'indexation à un document : les approches par extraction et les approches par assignation.

L'indexation automatique par extraction consiste à attribuer à un document des mots-clés qui sont extraits directement du document en question. Une approche classique de type statistique consiste à extraire des mots-clés d'un document en se basant sur leur degré de représentativité pour le document et sur leur degré de discrimination par rapport à l'ensemble des documents d'un corpus. C'est par exemple ce que permet le coefficient de pondération TF-IDF (Jones 1972). Selon ce coefficient, un mot d'un document sera sélectionné comme mot-clé s'il est à la fois très fréquent dans celui-ci et rare dans l'ensemble du corpus.

Certaines méthodes d'extraction reposent sur des algorithmes d'apprentissage automatique. Un exemple classique de méthode à base d'apprentissage automatique non-supervisé est l'indexation sémantique latente (LSI) (Deerwester et al. 1990) et un exemple important de méthode à base d'apprentissage automatique supervisé est l'algorithme KEA (Witten et al. 1999). De nombreux autres types de méthodes d'extraction existent, tout particulièrement des méthodes à base de graphes, qui centrent l'analyse sur la connexité des mots d'un document (Mihalcea and Tarau 2004).

L'indexation automatique par assignation consiste à aligner un document avec des mots-clés issus d'un vocabulaire contrôlé. Le défi particulier de cette approche est qu'elle implique d'attribuer à un document des mots-clés d'indexation qui ne sont pas nécessairement présents dans le texte de ce dernier. Cette approche a été l'objet d'importants travaux dans les dernières années, parmi lesquels on retrouve majoritairement des approches à base de graphes, par exemple la méthode TopicCoRank (Bougouin 2015) et des méthodes par apprentissage automatique supervisé, par exemple la méthode KEA++ (Medelyan et Witten 2008). Contrairement aux méthodes d'extraction qui consistent à sélectionner directement dans un document les mots saillants pour son indexation, les méthodes d'assignation opèrent en amont un travail d'analyse des équivalences entre les mots d'un document et un vocabulaire contrôlé. Ceci permet d'établir des relations de substitution. À la place de sélectionner comme mot-clé d'indexation un mot extrait d'un document, on sélectionne son ou ses meilleurs substituts dans un thésaurus.

La littérature dans le domaine tend à démontrer que l'indexation automatique, par extraction ou assignation, ne permet pas d'émuler de manière satisfaisante l'indexation manuelle des professionnels. Le développement de méthodes automatiques pour accomplir cette tâche représente toujours un important défi informatique.

La 12e édition du défi Fouille de Textes est une invitation à la communauté scientifique à apporter des éléments de réponses à la problématique de l'indexation automatique. Le reste de l'article présente les modalités du défi de la compétition, la méthode développée par notre équipe et l'analyse de ses performances.

2 Présentation du défi et des corpus

La tâche demandée dans le cadre de cette édition du défi est une tâche d'indexation automatique de notices bibliographiques de langue française. Ces notices sont regroupées en quatre corpus appartenant à quatre domaines de spécialité différente, soit la chimie (CHIMIE), les sciences de l'information (INFO), la linguistique (LING) et l'archéologie (ARCHEO). Le corpus CHIMIE contient 782 notices, le corpus INFO 706 notices, le corpus LING 715 notices et le corpus ARCHEO contient 718 notices. Chaque notice contient un document textuel (titre et résumé) et des mots-clés attribués manuellement par des indexeurs professionnels. Ces mots-clés peuvent être de type contrôlé (appartenant à un thésaurus) ou non-contrôlé.

Chaque corpus a été séparé de manière aléatoire en une base d'apprentissage et en base de test selon les proportions 2/3 et 1/3. Pour la compétition, les mots-clés des notices de la base de test de chaque corpus ont été supprimés. Le défi de la compétition consiste à développer une méthode qui permet de prédire automatiquement ces mots-clés. En d'autres mots, à partir de l'analyse des bases d'apprentissage, il faut prédire pour chaque document des notices des bases de test, quels sont les mots-clés attribués par les indexeurs professionnels.

3 Description de la méthode

La méthode développée par notre équipe est une méthode avec apprentissage automatique supervisée. La méthode est basée sur la construction de deux types d'espace sémantique. Le premier espace, dénoté ESMC, est un espace sémantique des mots-clés d'indexation. Cet espace est le modèle induit à partir de la base d'apprentissage. Le deuxième espace, dénoté ESD, est un espace sémantique des documents de la base de test. La prédiction des mots-clés pour un document de la base de test est réalisée en alignant les deux espaces ESMC et ESD. Les sous-sections suivantes présentent les différents paramètres de notre méthode.

3.1 L'espace sémantique des mots-clés d'indexation (ESMC)

L'ESMC modélise les relations de substitution ou d'équivalence entre d'une part les différentes chaînes de caractères qui composent les documents des notices d'une base d'apprentissage et d'autre part les mots-clés qui leur ont été attribués par des indexeurs professionnels. Dénotons par $\mathbb{T} = \{t_1 \dots t_n\}$ un ensemble de n mots-clés, par $\mathbb{C} = \{c_1 \dots c_m\}$ un ensemble de m chaînes de caractères et par θ_{ij} la valeur d'un coefficient d'association entre un mot-clé t_i et une chaîne c_j. Un ESMC peut être représenté par une matrice $[\theta_{ij}] \in \mathbb{R}^{n \times m}$ dans laquelle chaque ligne forme un vecteur $\vec{t}_i = (\theta_{i1} \dots \theta_{im})$ modélisant le degré d'équivalence entre le mot-clé t_i et l'ensemble des chaînes de caractères $c_j \in \mathbb{C}$.

Les mots-clés et les chaînes de caractères formant l'ESMC sont sélectionnés de la manière suivante. Les mots-clés $t_i \in \mathbb{T}$ sélectionnés correspondent à tous les mots-clés attribués par les indexeurs professionnels à un moins deux notices dans la base d'apprentissage d'un corpus. Ils incluent à la fois des mots-clés contrôlés membres d'un thésaurus et des mots-clés non-contrôlés. La sélection des chaînes de caractères $c_j \in \mathbb{C}$ est effectuée dans les titres et résumés des notices de la base d'apprentissage d'un corpus. Toutes les graphies sont premièrement récupérées. Ensuite, sont supprimées les graphies qui correspondent soit à des nombres, des singletons (un caractère) ou des mots fonctionnels. Les graphies retenues sont ensuite normalisées à l'aide d'un algorithme de racinisation. Finalement, les graphies racinées sont décomposées en chaînes de 5grams de caractères (Mcnamee and Mayfield 2004).

Différents coefficients d'association peuvent être utilisés pour calculer le degré d'équivalence θ_{ij} entre un mot-clé t_i et une chaîne c_j (e.g. information mutuelle, corrélation de Matthews, PMI). Ces coefficients doivent permettre de calculer la dépendance statistique entre t_i et c_j dans un corpus donné. Plus t_i est spécifique à c_j, plus la valeur de θ_{ij} doit être élevée. Dans nos expérimentations, le coefficient χ^2 (chi2) est celui qui a optimisé les performances de notre méthode (Manning, Raghavan, and Schütze 2008, 275).

3.2 L'espace sémantique des documents (ESD)

L'ESD modélise le contenu des documents en termes de distribution de chaînes de caractères. Dénotons par $\mathbb{D} = \{d_1 \dots d_n\}$ un ensemble de n document et par ω_{zj} la valeur d'un coefficient de pondération d'une chaîne de caractères $c_j \in \mathbb{C}$ dans le document $d_z \in \mathbb{D}$. Un ESD peut être représenté par une matrice $[\omega_{zj}] \in \mathbb{R}^{n \times m}$ dans lequel chaque ligne forme un vecteur $\vec{d}_z = (\omega_{z1} \dots \omega_{zm})$ modélisant la saillance des différentes chaînes de caractère $c_j \in \mathbb{C}$ dans le document d_z. L'ESD est forme essentiellement une variante basée sur des 5grams de caractères de l'espace vectoriel classique de Salton (Salton, Wong, and Yang 1975).

Cet espace est construit de la manière suivante. L'ensemble $\mathbb{D}$ correspond aux documents (titre et résumé) des notices de la base de test d'un corpus, alors que $\mathbb{C}$ représente l'ensemble des chaînes de caractères sélectionnés précédemment pour construire l'ESMC de telle sorte que l'ESMC et l'ESD d'un corpus partagent la même base vectorielle et sont alignés. Plusieurs coefficients de pondération peuvent être utilisés, la plupart étant des variantes du TF-IDF. Dans nos expérimentations, les performances de notre méthode ont été optimisées avec le coefficient Okapi BM25 (Robertson et al. 1999).

3.3 Prédiction des *k* mots-clés d'indexation

Notre méthode est basée sur l'hypothèse que la prédiction des mots-clés les plus pertinents pour indexer un document sont les mots-clés qui se substituent le mieux au contenu de ce document. La recherche de ces mots-clés est réalisée en comparant à l'aide d'une métrique les vecteurs $\vec{t}_i \in ESMC$ avec les vecteurs $\vec{d}_z \in ESD$. En d'autres mots, notre méthode est basée sur l'hypothèse que les k mots-clés les plus pertinents pour indexer le document d_z sont les k mots-clés dont les vecteurs $\vec{t}_i$ sont les plus similaires à $\vec{d}_z$. La prédiction consiste alors simplement à maximiser une métrique de similarité $\forall_{i=1}^{k} \underset{\vec{t}_i \in ESMC}{\operatorname{argmax}} S(\vec{d}_z, \vec{t}_i)$. Plusieurs métriques peuvent être utilisées pour comparer les vecteurs d'un espace sémantique (Kiela and Clark 2014, 23). Dans nos expérimentations, la métrique angulaire du cosinus est celle qui a optimisé les performances de notre méthode. Le nombre de k mots-clés prédit pour chaque document test a été fixé en utilisant le nombre moyen de mots-clés attribués par les indexeurs professionnels aux documents de la base d'apprentissage.

4 Résultats

La méthode présentée dans la section précédente est simple et construite à partir de techniques classiques qui ont fait leurs preuves dans de nombreux domaines de la recherche d'informations. Les résultats obtenus avec cette méthode sont toutefois parmi les meilleurs de la compétition. Sur les quatre tâches d'indexation automatique de la compétition, notre équipe a terminé première pour les tâches d'indexation des corpus INFO et LING, deuxième pour la tâche d'indexation du corpus CHIMIE et quatrième pour la tâche d'indexation du corpus ARCHEO. Le tableau 1 présente l'évaluation des performances de notre méthode ainsi que celle des méthodes des quatre autres équipes participantes. Les performances ont été évaluées selon la précision, le rappel et la F-mesure entre les mots-clés prédits par notre méthode et les mots-clés de référence attribués par les indexeurs professionnels. Les scores associés à une étoile sont ceux obtenus avec notre méthode.

Corpus	Évaluations	1er	2e	3e	4e	5e	moy.
INFO	Précision	***31.03**	21.26	21.92	11.72	13.83	19.952
	Rappel	***28.23**	30.32	21.83	23.54	12.01	23.186
	F-mesure	***28.98**	23.86	21.45	15.34	12.49	20.424
LING	Précision	***30.26**	23.28	23.16	13.98	15.67	21.27
	Rappel	***34.16**	32.73	25.85	30.81	16.1	27.93
	F-mesure	***31.75**	26.3	24.19	19.07	15.63	23.388
CHIMIE	Précision	24.92	***19.67**	21.15	10.88	18.19	18.962
	Rappel	21.73	***25.07**	17.54	30.25	14.9	21.898
	F-mesure	21.46	***21.07**	18.28	15.31	15.29	18.282
ARCHEO	Précision	43.48	55.26	53.77	***30.77**	33.93	43.442
	Rappel	52.71	38.03	33.46	***43.24**	31.25	39.738
	F-mesure	45.59	43.46	40.11	***34.96**	30.75	38.974

Tableau 1: Évaluation des performances des méthodes des participants de la compétition. Les scores associés à une étoile sont ceux obtenus avec notre méthode.

La performance de notre méthode sur les corpus INFO et LING dominent significativement celles des méthodes des autres participants. Sur le corpus CHIMIE, la performance de notre méthode est comparable à la méthode développée par l'équipe gagnante. Finalement, pour la tâche d'indexation automatique du corpus ARCHEO, la performance de notre méthode nous positionne en queue de peloton, bien que, en termes de performance absolue, elle soit la meilleure que nous ayons obtenue.

5 Discussion

Les résultats précédents sont caractérisés par d'importantes variations selon les corpus, ce qui suggère que notre méthode serait plus adaptée à certains types de corpus que d'autres. Les meilleures performances de notre méthode ont été obtenues sur le corpus ARCHEO, mais c'est sur les corpus INFO et LING que notre méthode se démarque le plus des autres. De plus, bien que notre méthode se positionne au deuxième rang pour la tâche d'indexation du corpus CHIMIE, les performances sont étonnamment faibles comparativement à celles obtenues sur les trois autres corpus.

Le tableau 2 présente quelques statistiques comparatives qui nous permettent d'expliquer ces résultats. La colonne « $\mathbb{C}_a \cap \mathbb{C}_t$ » correspond à la proportion de chaînes de caractères communes aux documents des notices d'apprentissage et de test. La colonne « $\mathbb{T}_a \cap \mathbb{T}_t$ » correspond à la proportion des mots-clés communes aux notices d'apprentissage et aux notices de test. La colonne « $\mathbb{T} \notin \mathbb{D}$ » correspond à la proportion de mots-clés attribué à une notice par les indexeurs professionnels qui ne sont pas occurrents dans le document (ni dans le titre ou le résumé). Finalement, la colonne « $\boldsymbol{I}(\mathbb{C}, \mathbb{T})$ » correspond à la quantité d'information mutuelle (Bouma 2009) entre les chaînes de caractères d'un corpus et ses mots-clés de références. Ce coefficient permet de mesurer, en termes de dépendance statistique, avec quelle régularité les indexeurs professionnels ont associé mots-clés et chaînes de caractères dans leur travail d'indexation. Plus la valeur de $\boldsymbol{I}(\mathbb{C}, \mathbb{T})$ est élevée, plus ce travail a été systématique : la présence de certaines chaînes de caractères dans un document est systématiquement associée à l'attribution de mots-clés spécifiques.

Corpus	F-mesure	Rang	$\mathbb{C}_a \cap \mathbb{C}_t$	$\mathbb{T}_a \cap \mathbb{T}_t$	$\mathbb{T} \notin \mathbb{D}$	$\boldsymbol{I}(\mathbb{C}, \mathbb{T})$
INFO	28.98	1er	0.67	0.55	0.68	7.15
LING	31.75	1er	0.71	0.58	0.61	8.33
CHIMIE	21.07	2e	0.70	0.44	0.76	3.82
ARCHEO	34.96	4e	0.73	0.60	0.37	10.10

Tableau 2 : Statistiques comparatives des corpus.

Ces statistiques comparatives nous permettent d'avancer quelques conjectures. Premièrement, les performances de notre méthode sont liées au degré de représentativité des bases d'apprentissage des corpus. C'est ce que montre la colonne « $\mathbb{T}_a \cap \mathbb{T}_t$ ». Plus les mots-clés de référence des notices de test sont absents des notices d'apprentissage, moins bonnes sont les prédictions réalisées par notre méthode. Ceci n'est pas étonnant puisqu'il s'agit d'une méthode avec apprentissage automatique de type supervisé, par conséquent la méthode ne peut prédire un mot-clé absent de la base d'apprentissage. Deuxièmement, la colonne « $\mathbb{T} \notin \mathbb{D}$ » montre que lorsqu'une proportion importante des mots-clés d'une notice sont également des mots qui figurent directement dans le document de la notice, comme c'est le cas pour le corpus ARCHEO, notre méthode performe moins bien que celles des autres équipes participantes. Troisièmement, les performances de notre méthode sont liées à la dépendance statistique entre mots-clés et chaînes de caractères. C'est ce que suggère la colonne « $\boldsymbol{I}(\mathbb{C}, \mathbb{T})$ ». Plus le travail des indexeurs professionnels sur un corpus a été systématique, c'est-à-dire plus l'association entre le contenu d'un document et ses mots-clés de référence est spécifique, plus les prédictions de notre méthode sont bonnes.

6 Conclusion

Les analyses précédentes suggèrent que notre méthode est particulièrement adaptée à des tâches d'indexation automatique qui nécessite une part importante d'assignation de mots-clés contrôlés. En effet, les performances de notre méthode se démarquent des celles des autres lorsque les mots-clés à prédire ne sont pas occurrents directement dans le document de la notice et par conséquent qu'une approche par extraction est peu pertinente. 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 performe moins bien sur le corpus ARCHEO. Nous conjecturons que les méthodes qui ont bien performé sur ce dernier corpus sont des approches qui exploitent davantage des techniques d'extraction de mots-clés. Les analyses précédentes suggèrent également que cet avantage de notre méthode sur les autres est toutefois conditionnel à la systématicité des indexations de référence. C'est pour cette raison que les performances de notre méthode sont moins dominantes pour la tâche d'indexation du corpus CHIMIE.

Références

Bougouin, Adrien. 2015. “Indexation Automatique Par Termes-Clés En Domaines de Spécialité.” Université de Nantes.

Bouma, Gerlof. 2009. “Normalized (pointwise) Mutual Information in Collocation Extraction.” *Proceedings of GSCL*, 31–40.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. “Indexing by Latent Semantic Analysis.” *Journal of the American Society for Information Science* 41 (6): 391–407.

Jones, Karen Spark. 1972. “A Statistical Interpretation of Term Specificity and Its Application in Retrieval.” *Journal of Documentation* 28 (1): 11–21.

Kiela, Douwe, and Stephen Clark. 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*, 21–30.

Manning, C., P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Mcnamee, Paul, and James Mayfield. 2004. “Character N-Gram Tokenization for European Language Text Retrieval.” *Information Retrieval* 7 (1-2): 73–97.

Mihalcea, Rada, and Paul Tarau. 2004. “TextRank: Bringing Order into Texts.” In . Association for Computational Linguistics.

Robertson, Stephen E., Steve Walker, Micheline Beaulieu, and Peter Willett. 1999. “Okapi at TREC-7: Automatic Ad Hoc, Filtering, VLC and Interactive Track.” *Nist Special Publication SP*, 253–64.

Salton, Gerard, Anita Wong, and Chung-Shu Yang. 1975. “A Vector Space Model for Automatic Indexing.” *Communications of the ACM* 18 (11): 613–20.

Witten, Ian H., Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig G. Nevill-Manning. 1999. “KEA: Practical Automatic Keyphrase Extraction.” In *Proceedings of the Fourth ACM Conference on Digital Libraries*, 254–55. ACM.

Indexation automatique de notices bibliographiques à l'aide d'approches d'acquisition terminologique

Thierry Hamon[1, 2]
(1) LIMSI, CNRS, Université Paris-Saclay, Bat 508, rue John von Neumann, Campus Universitaire, 91405 Orsay, France
(2) Université Paris 13 - Sorbonne Paris Cité, 99 avenue J.B. Clément, 93430 Villetaneuse, France
hamon@limsi.fr

RÉSUMÉ

Nous présentons dans cet article le système mis au point pour participer à la campagne DEFT 2016. Cette campagne ayant des objectifs similaires à celle de 2012, nous avons adapté le système utilisé pour DEFT 2012, afin de répondre aux contraintes de ce nouveau défi. Ainsi, les termes proposés par des approches d'acquisition terminologique sont regroupés en fonction de relations qu'ils entretiennent entre eux, puis sélectionnés à partir de leur position dans le texte et du vocabulaire qui les compose. Nous avons également tenté de prédire le nombre de mots-clés à l'aide d'un modèle de régression linéaire. Différentes configurations du système ont été appliquées à quatre domaines de spécialité. Les F-mesures des meilleures configurations varient entre 12,49 et 43,26.

ABSTRACT

Automatic indexing of bibliographic records with terminological acquisition approaches

In this paper, we present the system developed for the DEFT 2016 challenge. Since the challenge is similar to the challenge which took place in 2012, we adapted our DEFT 2012 system to satisfy the constraints of this new challenge. Thus, terms issued from terminological acquisition approaches are grouped according to relations between them. Then, these terms are selected according to their position in the text and the vocabulary they contain. We also attempt to predict the number of keyphrases per record with a linear regression model. Various configurations of the system have been used on the four specialized domains addressed in the challenge. The F-measures of the best configurations vary between 12.49 and 43.26.

MOTS-CLÉS : Mots-clés, acquisition terminologique, indexation contrôlée, extraction de termes.

KEYWORDS: Keyphrases, Terminological Acquisition, Controlled Indexing, Term Extraction.

1 Introduction

L'accès aux publications scientifiques à travers les notices bibliographiques s'appuie généralement sur une indexation réalisée par des professionnels de l'indexation documentaire. Avec l'augmentation du volume des publications [1] (Knoop *et al.*, 2015), l'indexation manuelle et l'assignation de mots-clés, issus ou non d'un thésaurus, deviennent des tâches de plus en plus complexes et coûteuses en temps.

1. https://www.nlm.nih.gov/bsd/medline_lang_distr.html, https://www.nlm.nih.gov/pubs/factsheets/medline.html

Il est donc nécessaire de mettre en place des systèmes qui aident les documentalistes dans leur travail et assurent une bonne qualité de l'indexation (Névéol *et al.*, 2007). Ce problème a conduit à de nombreuses propositions d'approches visant à extraire des mots ou des termes-clés à partir des résumés ou des articles scientifiques à indexer (voir (Kim *et al.*, 2010; Bougouin, 2013) pour un état de l'art), mais aussi à des campagnes comme SemEval 2010 [2] et DEFT 2012 [3]. C'est également la problématique à laquelle s'attaque cette nouvelle campagne DEFT à laquelle nous avons participé.

L'objectif est ici, de simuler l'indexation réalisée par un ingénieur documentaliste dans quatre domaines scientifiques : archéologie, chimie, linguistique, sciences de l'information. Il s'agit de réaliser une indexation contrôlée s'appuyant sur un thésaurus. Cette première indexation peut être complétée par des mots-clés issus ou non de la notice.

Après une présentation du matériel utilisé (section 2), nous décrirons l'approche mise en œuvre dans notre système à la section 3. Puis, nous présenterons et discuterons les résultats obtenus à la section 4.

2 Matériel

Lors de la phase d'entraînement, les participants à DEFT 2016 ont à leur disposition un corpus de notices bibliographiques et un thésaurus au format SKOS, pour chacun des domaines à traiter : archéologie, chimie, linguistique, sciences de l'information. Une notice est composée du titre et du résumé d'un article scientifique. La liste de mots-clés attribués par les documentalistes est également fournie avec chaque notice. Ces listes de mots-clés sont utilisée comme référence dans la campagne d'évaluation.

Chaque corpus d'entraînement est décomposé en un ensemble d'apprentissage et un ensemble de développement. Lors de la phase de test, des ensembles de notices issus des mêmes domaines étaient fournis. Les corpus de notices sont proposés dans trois formats : XML/TEI, texte brut et étiqueté morpho-syntaxiquement. Lors de nos expériences, nous avons utilisé les notices au format texte. Le tableau 1 présente le nombre de notices et le nombre de mots par domaine.

Domaine	Entraînement				Test	
	Apprentissage		Développement			
	notices	mots	notices	mots	notices	mots
archéologie	431	81 184	72	13 328	215	39 392
chimie	469	41 511	78	7 750	235	18 546
linguistique	429	56 041	71	9 472	215	29 648
sciences de l'information	424	39 456	70	7 070	212	26 304

TABLE 1 – Description des corpus d'entraînement et de test (nombre de documents et de mots).

Pour chaque domaine, les organisateurs du défi ont fourni le thésaurus utilisé lors de l'indexation manuelle et dont une partie des mots-clés associés aux notices est issue (tableau 2). Les thésaurus proposent des termes préférés (URI `prefLabel` ; par ex. *Alfacalcidol*) et des variantes (URI `altLabel` ; par ex. *1-α-Hydroxycolecalciferol* et *1-α-Hydroxyvitamin D3*). L'analyse des thésaurus montre d'une part que la proportion de variantes terminologique est variable suivant les domaines :

2. `http://semeval2.fbk.eu/semeval2.php`
3. `https://deft.limsi.fr/2012/`

entre 5,2 % dans le thésaurus d'archéologie et 16,7 % dans le thésaurus de sciences de l'information. D'autre part, la taille et le contenu des thésaurus varient également : si les thésaurus des domaines de l'archéologie et de la linguistique sont relativement petits et couvrent spécifiquement leurs domaines respectifs, les thésaurus des domaines de la chimie et des sciences de l'information sont multi-domaines et partagent beaucoup de termes entre eux : les termes du thésaurus de sciences de l'information se trouvent tous dans le thésaurus de chimie. Cette situation constitue une difficulté particulière pour le défi.

Domaine	Termes préférés	Variantes Terminologiques	Total
archéologie	4 339	238 (5,2 %)	4 577
chimie	124 762	19 992 (13,8 %)	144 754
linguistique	13 052	1 214 (8,5 %)	14 266
sciences de l'information	103 225	20 632 (16,7 %)	123 857

TABLE 2 – Description des thésaurus par domaine.

3 Méthode et description du système

Étant donné les similarités entre les défis DEFT 2012 et DEFT2016, nous avons choisi d'utiliser le système que nous avions mis au point pour DEFT 2012 (Hamon, 2012). Toutefois, nous l'avons adapté pour répondre aux contraintes de ce nouveau défi. Ainsi, une première étape s'appuie sur des méthodes d'acquisition terminologique (section 3.1). Les termes extraits sont ensuite pondérés, filtrés et sélectionnés (section 3.2). Contrairement à DEFT 2012, le nombre de mots-clés par notice n'est pas fourni. Pour répondre à ce problème, nous avons mis au point plusieurs stratégies (section 3.3).

3.1 Acquisition terminologique

L'identification des mots-clés des notices est principalement réalisée grâce à une reconnaissance des termes du thésaurus. Pour cela, ces termes contrôlés ont été projetés sur les titres et les résumés à l'aide de TermTagger [4]. Les textes des notices et les termes issus d'un thésaurus sont lemmatisés grâce à TreeTagger (Schmid, 1997) pour augmenter la couverture de la reconnaissance. Outre la prise en compte de la casse ou de variations typographiques (par exemple, présence d'un trait d'union à la place d'un espace ou inversement), nous avons fait évolué l'outil afin de tenir compte l'absence de majuscules accentuées dans les termes des thésaurus. Sur des expériences préliminaires, nous avons constaté que cette modification apporte un gain de 0,5 à 0,75 de F-mesure.

Afin d'étendre la couverture de la reconnaissance des mots-clés, nous avons complété les résultats obtenus en utilisant, d'une part, les termes simples et complexes extraits automatiquement par YaTeA [5] (Aubin & Hamon, 2006) et, d'autre part, des variantes morpho-syntaxiques des termes du thésaurus fournies par Faster (Jacquemin, 1997).

4. http://search.cpan.org/~thhamon/Alvis-TermTagger/
5. http://search.cpan.org/~thhamon/Lingua-YaTeA/

3.2 Pondération et filtrage des termes

Nous avons tout d'abord filtré les listes de mots-clés candidats selon deux critères : (i) les termes étiquetés comme adjectifs ne sont pas conservés, (ii) les termes extraits automatiquement par YATEA qui contiennent au moins un mot plein issu de la liste des mots clés sont retenus. Ainsi, de manière similaire à (Drouin, 2003), nous faisons l'hypothèse que si les mots-clés sont constitués de mots caractéristiques du domaine.

Bien que l'évaluation ne nécessite pas d'ordonner les termes obtenus lors de la première étape, les stratégies que nous avons utilisées pour sélectionner les termes nous ont conduit à leur associer un poids devant refléter leur pertinence par rapport à la tâche. Ainsi, nous avons surtout considéré la position de la première occurrence du terme, c'est-à-dire le nombre de caractères depuis le début du résumé. Nous faisons l'hypothèse que les termes situés au début du résumé ont un poids plus élevé que ceux situés vers la fin du résumé. Les termes sont alors triés dans l'ordre décroissant. Nous avons également considéré le nombre d'occurrences des formes canoniques des termes après regroupement en fonction des lemmes et des relations de variation morpho-syntaxique, le nombre d'occurrences d'une forme canonique étant la somme du nombre d'occurrences des formes fléchies de termes.

3.3 Détermination du nombre de mots-clés et sélection

La dernière étape de notre approche consiste à réduire la liste de mots-clés candidats. Notons que cette sélection est particulièrement nécessaire lorsque les termes extraits par YATEA sont intégrés à la liste.

Nous avons donc envisagé trois stratégies pour déterminer le nombre de mots-clés associés à chaque notice : (i) le nombre moyen de mots-clés par notice pour un domaine donné, calculé sur l'ensemble d'apprentissage (17 pour l'archéologie, 13 pour la chimie, 10 pour la linguistique, et 9 pour les sciences de l'information) (ii) la prédiction du nombre de mots-clés pour chaque notice à l'aide d'un modèle de régression linéaire (Yan & Su, 2009), appris sur le nombre de mots-clés des notices de l'ensemble d'apprentissage de chaque domaine, les traits étant le nombre de mots, de phrases, de noms, d'adjectifs et de verbes de la notice, (iii) utilisation de la liste complète sans limitation du nombre de mots-clés. L'objectif de la deuxième stratégie est de réduire l'écart entre le nombre attendu de mots-clés et le nombre de mots-clés sélectionnés par une méthode automatique.

3.4 Configuration du système

L'ensemble des traitements a été réalisé dans la plate-forme Ogmios (Hamon & Nazarenko, 2008). Chaque notice a été segmentée en mots et en phrases. Nous avons utilisé le TreeTagger (Schmid, 1997) pour l'étiquetage morpho-syntaxique et la lemmatisation des mots des notices et des termes des thésaurus. Nous avons effectué plusieurs expériences afin d'identifier les combinaisons de paramètres les plus adaptées pour l'identification des mots-clés.

Ainsi, la première méthode d'identification des mots-clés d'une notice (**m1**) consiste à reconnaître, à l'aide de TermTagger, les termes issus des thésaurus des domaines mais aussi les mots-clés associés aux notices des ensembles d'entraînement. Lorsqu'aucun terme de thésaurus n'est reconnu dans une

notice [6], tous les termes extraits par YATEA sont utilisés. Les variantes morpho-syntaxiques des termes des thésaurus sont acquis sur les notices avec Faster (Jacquemin, 1997) sur la base des termes extraits par YATEA. Les mots-clés candidats sont ordonnés selon la position de leur première occurrence. Les filtres décrits à la section 3.2 sont ensuite appliqués. Tous les mots-clés sont associés aux notices.

La deuxième méthode (**m2**) est similaire à la précédente. Cependant, les variantes morpho-syntaxiques ne sont pas prises en compte et nous sélectionnons les n premiers mots-clés de la liste, où n est le nombre moyen de mots-clés par notice pour chaque domaine.

Pour la troisième méthode, Nous avons choisi d'utiliser des méthodes différentes selon les domaines étant donné sur les résultats d'expériences préliminaires réalisées sur les corpus d'entraînement. Ces expériences visaient à déterminer les configurations les plus adaptées selon les domaines. Ainsi, pour les notices de linguistique, les mots-clés candidats sont triés selon le nombre d'occurrences des formes canoniques (**m3b**) tandis qu'en sciences de l'information, les termes extraits par YATEA ont été intégrés à la liste des mots-clés indépendamment des résultats de la projection du thesaurus sur les notices (**m3c**). Pour les quatre domaines, le nombre de mots-clés sélectionnés est déterminé à l'aide du modèle de régression linéaire. Pour les notices d'archéologie et de chimie, l'identification des mots-clés à partir des notices est identique à celle utilisée dans la méthode **m1**. Seule la détermination du nombre de mots-clés associés aux notices change. La méthode est notée **m3a**.

4 Résultats et discussion

Les résultats obtenus sur les corpus de test sont présentés dans le tableau 3. A l'exception du domaine de l'archéologie, où la troisième méthode (**m3**) est la plus performante, les meilleures F-mesures sont obtenues avec la deuxième méthode (**m2**). Nous observons également que, quel que soit le domaine, la première méthode (**m1**) permet d'obtenir le rappel le plus élevé. Par contre, la meilleure précision est en général obtenue avec la deuxième méthode (**m2**), sauf pour les notices de chimie pour lesquelles la meilleure précision est obtenue avec la troisième méthode d'indexation (**m3**).

En comparaison avec la méthode **m2**, le modèle de régression linéaire (méthode **m3**) visant à prédire le nombre de mots-clés dégrade assez peu les résultats sur les notices d'archéologie et de chimie en terme de F-mesure.

Nous observons également que les résultats sont bien meilleurs pour le domaine de l'archéologie que pour les autres domaines. Cela pourrait s'expliquer par les pratiques des documentalistes qui indexent les notices en s'appuyant davantage sur le thésaurus, mais aussi, *a contrario*, par l'utilisation de thésaurus multi-domaines en chimie et en sciences de l'information pour l'indexation automatique ainsi que par l'ambiguïté des termes utilisés dans les domaines et par le degré de consensus sur la terminologique du domaine. Enfin, bien que le contexte et les données ne soient pas identiques (les articles de DEFT 2012 étant issus de revues de Sciences Humaines et Sociales et la liste des mots-clés associés à tous les articles était fournie), les résultats obtenus sur les notices de linguistique sont plus faibles que ceux de DEFT 2012 : la F-mesure variait entre 0,2253 et 0,3985.

6. C'est le cas pour une notice bibliographique en chimie.

Domaine/Méthode	Précision	Rappel	F-mesure
archéologie			
m1	54,35	38,51	42,92
m2	55,30	37,92	43,19
m3a	55,26	38,03	**43,26**
chimie			
m1	17,26	15,83	15,14
m2	18,19	14,90	**15,29**
m3a	18,30	14,77	15,27
linguistique			
m1	13,88	18,75	15,36
m2	15,67	16.10	**15.63**
m3b	15.12	14.76	14.71
sciences de l'information			
m1	11,75	14,42	12,27
m2	13,83	12,01	**12,49**
m3c	12,28	11,58	11,56

TABLE 3 – Résultats sur le corpus de test.

5 Conclusion

Afin d'identifier des mots-clés de notices bibliographiques, nous avons mis en œuvre des méthodes d'acquisition terminologique. Les termes fournis sont ordonnés en fonction de leur position dans la notice puis sélectionnés en fonction du vocabulaire qui les compose. Nous avons également cherché à prédire le nombre de mots-clés à l'aide d'un modèle de régression linéaire.

Il est également à noter que la précision, le rappel et la F-mesure ayant été utilisés pour évaluer les systèmes, le nombre de mots-clés a une influence sur les résultats. De plus, on considère ici la tâche comme complètement automatisée, alors qu'elle peut être partiellement subjective (Bougouin *et al.*, 2016). La MAP (moyenne des précisions moyennes) aurait pu permettre de se placer davantage dans une perspective d'aide à l'identification des mots-clés puisqu'elle tient compte de l'ensemble des mots-clés pertinents dans la liste et de leur rang.

En perspective de ce travail, nous souhaitons prendre en compte les variantes sémantiques des termes comme les synonymes et les hyperonymes, mais aussi de les utiliser lors du regroupement des termes sous une même forme canonique. De plus, dans une perspective d'indexation totalement automatique, la prédiction du nombre de mots-clés permettant d'obtenir de résultats assez proches de ceux basés sur le nombre moyen de mots-clés par notice, nous envisageons d'utiliser d'autres traits pour améliorer le modèle.

Références

AUBIN S. & HAMON T. (2006). Improving term extraction with terminological resources. In T. SALAKOSKI, F. GINTER, S. PYYSALO & T. PAHIKKALA, Eds., *Advances in Natural Language*

Processing (5th International Conference on NLP, FinTAL 2006), number 4139 in LNAI, p. 380–387 : Springer.

BOUGOUIN A. (2013). État de l'art des méthodes d'extraction automatique de termes-clés. In *Actes de la 15e Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL'2013)*, p. 96–109, Les Sables d'Olonne, France.

BOUGOUIN A., BARREAUX S., ROMARY L., BOUDIN F. & DAILLE B. (2016). Termith-eval : a french standard-based resource for keyphrase extraction evaluation. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, M. GROBELNIK, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Eds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia : European Language Resources Association (ELRA).

DROUIN P. (2003). Acquisition automatique des termes simples fondée sur les pivots lexicaux spécialisés. In *Actes de la Conférence TIA-2003*, p. 183–186, Strasbourg. Poster.

HAMON T. (2012). Acquisition terminologique pour identifier les mots clés d'articles scientifiques. In *Actes de l'atelier DEFT 2012*, p. 25–31, Grenoble, France.

HAMON T. & NAZARENKO A. (2008). Le développement d'une plate-forme pour l'annotation spécialisée de documents web : retour d'expérience. *Traitement Automatique des Langues*, **49**(2), 127–154.

JACQUEMIN C. (1997). *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus*. Mémoire d'habilitation à diriger des recherches en informatique fondamentale, Université de Nantes.

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, Uppsala, Sweden : Association for Computational Linguistics.

KNOOP M., BRIGHIGNI L. & VAN TIGGELEN B. (2015). Les publications en physique au cnrs : combien, par qui et où ? *Reflets de la Physique*, **43**, 58–60.

NÉVÉOL A., SHOOSHAN S. E., HUMPHREY S. M., RINDFLESH T. C. & ARONSON A. R. (2007). Multiple approaches to fine-grained indexing of the biomedical literature. In *Pacific Symposium on Biocomputing*, volume 12, p. 292–303.

SCHMID H. (1997). Probabilistic part-of-speech tagging using decision trees. In D. JONES & H. SOMERS, Eds., *New Methods in Language Processing Studies in Computational Linguistics*.

YAN X. & SU X. G. (2009). *Linear Regression Analysis : Theory and Computing*. River Edge, NJ, USA : World Scientific Publishing Co., Inc.

LIPN@DEFT2016 : Annotation de documents en utilisant l'Information Mutuelle

Davide Buscaldi Haïfa Zargayouna
Laboratoire d'Informatique de Paris Nord (LIPN, UMR 7030)
Université Paris 13 – Sorbonne Paris Cité & CNRS
{prenom.nom}@lipn.univ-paris13.fr

RÉSUMÉ

Cet article décrit le système proposé par le LIPN pour participer à l'édition 2016 pour le Défi Fouille de Textes. La tâche consiste à proposer des mots clés pour indexer des notices bibliographiques. Quatre domaines de spécialités ont été proposés : linguistique, sciences de l'information, archéologie et chimie. Nous avons proposé trois approches : une approche qui s'appuie sur le volet terminologique des thesaurus, une approche fondée sur l'information mutuelle et une approche qui fusionne les deux. Les mêmes approches ont été appliquées aux quatre domaines de spécialité.

ABSTRACT

LIPN@DEFT2016 : Document annotation based on pointwise Mutual Information

This paper describes the participation of the LIPN team at DEFT 2016. The task consisted on generating keywords in order to index bibliographic records. We proposed three approaches to solve the annotation problem in all the four domains offered (linguistics, information science, archeology and chemistry). The first approach is based on the terminology that is contained in domain ontologies or thesauri, the second one is based on pointwise mutual information, while the third on is a combination of the two approaches.

MOTS-CLÉS : Annotation Sémantique, Information Mutuelle, Fouille de Texte.

KEYWORDS: Semantic Annotation, Mutual Information, Text Mining.

1 Introduction

L'édition 2016 du défi a consisté à proposer des mots clés pour indexer une notice bibliographique. Ces mots clés peuvent être contrôlés ou libres. Des thesaurus sont fournis par domaine. Ils permettent de sélectionner les descripteurs censés caractériser au mieux le contenu des notices. Quatre domaines de spécialité sont proposés : linguistique, sciences de l'information, archéologie et chimie. Quand on dispose de thesaurus, la tâche consiste à proposer une indexation ou une annotation sémantique.

Dans nôtre participation à DEFT, nous avons identifié deux cas dans lesquels un processus d'annotation sémantique peut être mis en place :

1. l'étiquette du concept est dans le texte ;
2. l'étiquette n'est pas dans le texte mais il y a des indices repérables dans le texte.

Dans le premier cas, nous avons choisi de chercher directement dans le texte les étiquettes du thesaurus avec un moteur de recherche. Dans le deuxième cas, nous avons choisi d'exploiter le corpus

d'entraînement pour extraire les termes qui s'associent le plus souvent à certains descripteurs ou concepts. Finalement cela correspond à calculer l'information mutuelle entre les concepts et les termes des textes annotés dans le corpus d'entraînement. Nous ne nous sommes pas intéressés aux cas où les descripteurs ont été librement proposés par les documentalistes.

Dans la littérature, plusieurs approches d'annotation automatique existent. Parmi les plus efficaces, on peut citer une méthode de classification fondée sur les K plus proches voisins appliquée au domaine médical par exemple (Trieschnigg *et al.*, 2009), et dans le même domaine, la proposition de classificateurs bayesiens et machines à vecteurs de support (SVM) par (Jimeno Yepes *et al.*, 2012). L'information mutuelle a aussi été utilisée efficacement pour sélectionner des caractéristiques dans le contexte de l'annotation ou catégorisation de texte, individuellement (Lewis, 1992) ou en conjonction avec des algorithmes d'apprentissage supervisés (Dumais, 1998; Wang & Lochovsky, 2004). Dans notre proposition, l'information mutuelle est utilisée directement, sans passer par l'intermédiaire d'un algorithme d'apprentissage. Il est en effet, difficile voire impossible d'envisager de proposer un modèle par concept.

L'annotation fondée sur le volet terminologique d'une ontologie nécessite de mettre en place des techniques de reconnaissance de termes dans les textes comme par exemple dans (Jacquemin, 1997). L'approche que nous proposons repose sur un moteur de recherche. Elle ne nécessite aucune analyse fine des textes et permet de faire une reconnaissance approximée des termes. Nous décrivons dans la section suivante les algorithmes d'annotation que nous avons testés. Dans la section 3, nous présentons les résultats obtenus sur les différents corpus.

2 Algorithmes d'annotation

Nous présentons brièvement les deux techniques d'annotations proposées. La première technique s'appuie sur les étiquettes des concepts, la deuxième technique s'appuie sur la co-occurence entre les termes et les concepts.

2.1 Annotation fondée sur la terminologie

Cet algorithme d'annotation est utilisé par le moteur de recherche d'information sémantique YaSemIR (*Yet another Semantic Information Retrieval system*) (Buscaldi & Zargayouna, 2013) [1]. D'abord, la ressource sémantique (ontologie ou thesaurus) est analysée et indexée. L'index crée, relie chaque identifiant de concept à ses représentations lexicales ou étiquettes (principales ou autres) qui constituent le volet terminologique de la ressource. Les fragments textuels qui constituent les étiquettes peuvent ainsi être cherchés par le moteur de recherche.

En phase d'annotation, le texte du document à annoter d est analysé pour extraire les fragments à soumettre au moteur de recherche en tant que requêtes. Annoter un document revient à interroger la ressource indexée. Dans YaSemIR, seuls les syntagmes nominaux sont pris en compte. L'analyse syntaxique de tous les documents s'avérant très lente, pour DEFT2016 nous avons choisi d'extraire les trigrammes. Le moteur de recherche renvoie une liste ordonnée de résultats contenant les IDs des concepts avec leurs poids. Le meilleur résultat est ajouté à l'ensemble d'annotations du documents. La Figure 1 montre un exemple d'annotation en utilisant l'ontologie du domaine archéologique.

1. https://github.com/dbuscaldi/YaSemIR

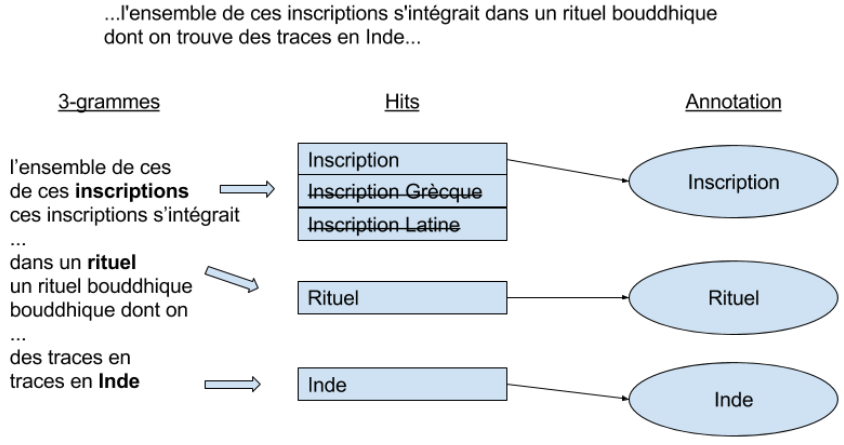

FIGURE 1 – Exemple de chaîne de traitement de l'annotateur YaSemIR

Cet algorithme d'annotation est utile quand les étiquettes sont présentes dans le texte même partiellement. Le bruit peut provenir des étiquettes partageant la même racine ainsi que des concepts retournés en premier mais avec un faible score.

2.2 Annotation fondée sur l'Information Mutuelle

Quand les concepts n'ont pas de représentations lexicales dans le texte, les algorithmes d'annotation fondés sur la terminologie échouent. Nous proposons un algorithme qui se fonde sur la co-occurrence entre le concept et les termes du documents. L'idée est que si, dans un corpus de textes annotés, un ou plusieurs termes co-occurrent souvent avec le même concept c, alors si on retrouve le même terme ou la même combinaison de termes dans un document sans annotation, on peut l'annoter avec le concept c.

Nous avons d'abord procédé à quelques filtrages. Nous avons éliminé tous les cas pour lesquels les co-occurrences sont dues au pur hasard. Pour cela, dans la phase d'entraînement nous avons pris en compte seulement les concepts qui ont servi pour annoter au moins 5 documents dans le corpus de référence. Nous n'avons pris en compte que les substantifs et adjectifs qui représentent habituellement les étiquettes des concepts.Au final, nous avons calculé l'information mutuelle uniquement si le terme et le concept co-occurrent dans au moins 3 documents.

L'information mutuelle ponctuelle est définie comme :

$$IP(t,c) = \log \frac{p(t,c)}{p(t)p(c)}$$

Où t est un terme (substantif ou adjectif dans notre cas) et c un concept ; $p(t,c)$ est la probabilité conjointe entre t et c, calculée comme $freq(t,c)/N$, $p(t) = freq(t)/N$ et $p(c) = freq(c)/N$, où N est le nombre de documents dans le corpus d'entraînement.

2.2.1 Construction du modèle

Nous construisons une matrice M avec $|C|$ lignes et $|T|$ colonnes, où C est l'ensemble des étiquettes dans le corpus d'entraînement avec fréquence > 5, et T est l'ensemble des mots du dictionnaire (donc tous les substantifs et adjectifs observés dans le corpus) avec fréquence > 3. Chaque élément $M_{i,j}$ est calculé comme :

$$M_{i,j} = \begin{cases} IP(t_i, c_j) & \text{if } IP(t_i, c_j) > 0 \\ 0 & \text{else} \end{cases}$$

Sur cette matrice, nous appliquons l'analyse sémantique latente (LSA - *Latent Semantic Analysis*), en décomposant M en valeurs singulières (algorithme SVD) $M = U\Sigma V^T$, et en approximant M avec $\hat{M} = U_k \Sigma_k V_k^T$, avec les meilleurs k valeurs singulières sélectionnés ($k = 100$).

Nous espérons améliorer la couverture en appliquant LSA sur la matrice. En effet, LSA permet de trouver des termes fortement associés avec d'autres termes qui sont très caractéristiques pour certaines étiquettes, même si dans le corpus d'entraînement l'étiquette n'apparaît pas avec ce mot. Par exemple, si « Paris » et « français » co-occurrent souvent, mais dans la collection on ne trouve que « France » associé au terme « français », on peut déduire que « France » est une annotation plausible si dans le texte on trouve « Paris », même si on ne les a jamais vus ensemble dans le corpus l'entraînement.

Finalement, nous appliquons un filtre sur la matrice $\hat{M}$ de la façon suivante :

$$\hat{M}^{\sigma}_{i,j} = \begin{cases} \hat{M}_{i,j} & \text{if } \hat{M}_{i,j} \geq \left(0.5 * \sigma'(\hat{M}_{*,j}) + \mu'(\hat{M}_{*,j})\right) \\ 0 & \text{else} \end{cases}$$

Où $\hat{M}_{*,j}$ est le $j-$ème vecteur colonne de la matrice $\hat{M}$, $\sigma'(\mathbf{x})$ est l'écart type des éléments non nuls du vecteur $\mathbf{x}$ et $\mu'(\mathbf{x})$ est la moyenne des éléments non nuls du vecteur $\mathbf{x}$. La matrice $\hat{M}^{\sigma}$ ainsi obtenue est utilisée comme *modèle* pour l'annotation des documents.

Nous avons éliminé du modèle les termes non discriminants afin d'améliorer la précision.

2.2.2 Annotation

Nous associons au document d à annoter un vecteur binaire $\mathbf{b}$ de taille $|T|$ (taille du vocabulaire) où chaque élément $\mathbf{b}_i$ est à 1 si le terme correspondant est dans le texte du document d, 0 autrement. Nous calculons pour chaque étiquette l_i le score $s(l_i) = \mathbf{b} \cdot \hat{M}^{\sigma}_{i,*}$. Le document est annoté finalement avec les 20 étiquettes avec les meilleurs scores > 0. S'il y a moins de 20 étiquettes avec des scores positifs, alors une annotation est générée avec toutes les étiquettes ayant un score positif. Le choix de 20 a été défini empiriquement après les tests sur le corpus de développement.

L'exemple en figure 2 montre un cas d'annotation, avec une bonne étiquette (« Recherche scientifique ») et une mauvaise (« système de recherche d'information »). Nous pouvons observer que le modèle ne prend pas en compte l'absence de caractéristiques importantes dans les vecteurs (le terme « système »). Un algorithme d'apprentissage pourrait tirer profit de ces informations.

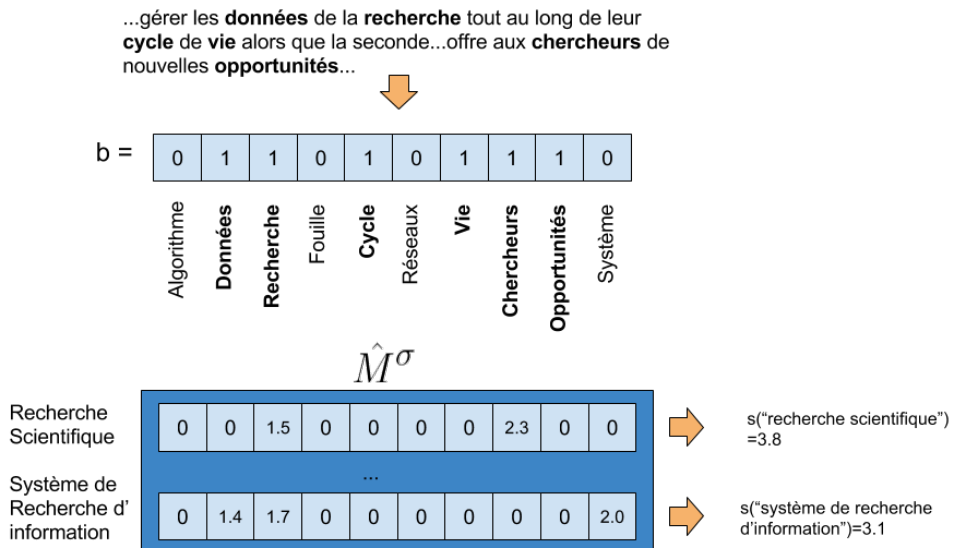

FIGURE 2 – Exemple d'annotation en utilisant l'IM sur un document du domaine *sciences de l'information*. Les termes en gras sont les termes du texte à annoter présents dans le thesaurus

3 Résultats

Les modèles utilisés pour la participation à DEFT2016 ont été les suivants : modèle M1 - annotateur YaSemIR ; modèle M2 - annotateur fondé sur l'information mutuelle ; modèle M3 - fusion des résultats M1 et M2. L'idée du troisième modèle est d'exploiter la complémentarité entre M1 et M2. Ci dessous les résultats obtenus sur le corpus d'entraînement et développement avec M1 et M2 respectivement.

Corpus	Prec	Rappel	F-Measure
archeo	19.61%	18.86%	19.23%
chimie	6.44%	9.61%	7.71%
linguistique	1.08%	2.26%	1.46%
sciences info	6.13%	14.16%	8.56%

TABLE 1 – Résultats sur le corpus d'entraînement avec YaSemIR, sur les différents corpus.

La comparaison des résultats (voir tableau 2) montre clairement l'avantage obtenu grâce à l'application de LSA sur la matrice M. Les résultats officiels (Tableau 3) montrent que la fusion a permis de profiter effectivement de la complémentarité des sorties pour améliorer le rappel, mais la faible précision des méthodes a empêché d'obtenir des améliorations par rapport à la F-mesure, sauf dans le cas du domaine de la chimie.

Corpus	LSA	Prec	Rappel	F-Measure
archeo	n	16.22%	20.50%	18.11%
	y	20.71%	26.67%	23.31%
chimie	n	11.62%	13.68%	12.56%
	y	12.53%	20.16%	15.45%
linguistique	n	2.95%	7.17%	4.18%
	y	5.75%	13.66%	8.09%
sciences info	n	12.60%	21.62%	15.92%
	y	13.07%	28.68%	17.95%

TABLE 2 – Résultats sur le corpus de développement avec IM - avec et sans LSA.

Corpus	Modèle	Prec	Rappel	F-Measure
archeo	M1	**33.93%**	31.25%	**30.75%**
	M2	19.67%	24.01%	20.95%
	M3	22.63%	**46.54%**	29.61%
chimie	M1	**13.87%**	16.00%	13.29%
	M2	11.13%	17.76%	13.08%
	M3	10.88%	**30.25%**	**15.31%**
linguistique	M1	10.04%	10.45%	12.97%
	M2	**13.98%**	30.81%	**19.07%**
	M3	10.59%	**45.23%**	16.98%
sciences info	M1	8.86%	18.65%	11.48%
	M2	**11.72%**	23.54%	**15.34%**
	M3	9.03%	**36.61%**	14.25%

TABLE 3 – Résultats officiels sur le corpus d'évaluation avec les modèles M1, M2 et la fusion des sorties (M3).

4 Conclusions

Les deux difficultés majeures de la tâche proposée à DEFT2016 était le grand nombre d'étiquettes et le fait que les étiquettes pouvaient provenir ou pas d'un thésaurus. Nous avons choisi d'utiliser un annotateur fondé sur la terminologie pour tirer profit au mieux des thesauri. Vu la difficulté d'entraîner un modèle pour des milliers d'étiquettes, nous avons choisi d'utiliser un algorithme d'annotation fondé sur l'information mutuelle. Ce dernier a été mis au point sur le corpus de développement, mais les différents paramètres qui ont été choisis (fréquence minimale des étiquettes, poids des termes dans les vecteurs, nombre d'étiquettes à retourner, en particulier) n'ont pas été testés de façon approfondie. Nous estimons donc qu'il reste une marge d'amélioration pour cette méthode, cela pourra être vérifié par des expériences plus étendues. Il est aussi à noter que pour bien estimer l'information mutuelle, nous avons besoin d'un corpus d'entraînement assez grand. Dans le corpus d'entraînement DEFT2016, nous estimons qu'il n'y a assez de données que pour 12.47% des étiquettes en moyenne ; par exemple pour les 2256 étiquettes trouvées dans le corpus apprentissage pour le domaine de l'archéologie, seulement 278 ont une fréquence supérieure à 5. Nous pourrons augmenter le seuil de fréquence pour améliorer la précision du modèle mais il faudrait compenser la perte de rappel. Finalement, cette expérience nous a permis d'observer l'utilité du LSA dans ce contexte.

Remerciements

Cette recherche s'insère dans le programme « Investissements d'Avenir » géré par l'Agence Nationale de la Recherche ANR-10-LABX-0083 (Labex EFL).

Références

BUSCALDI D. & ZARGAYOUNA H. (2013). YaSemIR : Yet Another Semantic Information Retrieval System. In *Proceedings of the Sixth International Workshop on Exploiting Semantic Annotations in Information Retrieval*, ESAIR '13, p. 13–16, New York, NY, USA : ACM.

DUMAIS S. (1998). Using SVMs for text categorization. *IEEE Intelligent Systems*, **13**(4), 21–23.

JACQUEMIN C. (1997). *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus.* Habilitation à diriger des recherchesinformatique fondamentale, Université de Nantes.

JIMENO YEPES A., MORK J. G., WILKOWSKI B., DEMNER FUSHMAN D. & ARONSON A. R. (2012). MEDLINE MeSH Indexing : Lessons Learned from Machine Learning and Future Directions. In *Proceedings of the 2Nd ACM SIGHIT International Health Informatics Symposium*, IHI '12, p. 737–742, New York, NY, USA : ACM.

LEWIS D. D. (1992). Feature Selection and Feature Extraction for Text Categorization. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '91, p. 212–217, Stroudsburg, PA, USA : Association for Computational Linguistics.

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), 1412–1418.

WANG G. & LOCHOVSKY F. H. (2004). Feature Selection with Conditional Mutual Information Maximin in Text Categorization. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM '04, p. 342–349, New York, NY, USA : ACM.

Représentation vectorielle de mots pour l'indexation de notices bibliographiques

Morgane Marchand[1] Geoffroy Fouquier[1] Guillaume Pitel[1]
(1) eXenSa, 41 rue Périer, 92120 Montrouge
morgane.marchand@exensa.com, geoffroy.fouquier@exensa.com,
guillaume.pitel@exensa.com

RÉSUMÉ

Cet article présente la contribution d'eXenSa à l'édition 2016 au DÉfi Fouille de Textes (DEFT) dont la tâche consiste à indexer des documents scientifiques par des mots clefs, préalablement sélectionnés par des professionnels. Le système proposé est purement statistique et combine une approche graphique et une approche sémantique. La première approche cherche dans le titre et le résumé du document des mots graphiquement proches des mots clefs du thésaurus. La seconde approche attribue à un nouveau document des mots clefs associés aux documents du corpus d'entraînement qui lui sont sémantiquement proches. Les deux approches utilisent des représentations vectorielles apprisent en utilisant l'algorithme NCISC, un algorithme stochastique de factorisation de matrices. Notre approche s'est classée première sur deux des corpus de spécialité proposés et deuxième sur les deux autres.

ABSTRACT

Word vector embeddings for bibliographic records indexing

This article presents the eXenSa contribution to the 2016 DEFT Workshop. The proposed task consists in indexing bibliographic records with keywords chosen by professional indexers. We propose a statistical approach which combines graphical and semantic approches. The first approach defines a document keywords as thesaurus terms graphically similars to terms contained in the title or the abstract of this document. The second approach assigns to document the keywords associated with semantically similar documents in training corpora. Both approach use models generated using NCISC, a stochastic matrix factorisation algorithm. Our system obtains the best F-score on half of the four test corpuses and ranks second for the two others.

MOTS-CLÉS : Indexation, mots clefs, espaces sémantiques, représentation vectorielle de mots.

KEYWORDS: Indexation, keywords, semantic spaces, word vector embedding.

1 Introduction

L'édition 2016 de DEFT propose de travailler sur l'indexation de documents scientifiques par des mots clefs qui ont été proposés par des indexeurs professionnels. Les corpus proposés sont issus de quatre spécialités : linguistique, sciences de l'information, archéologie et chimie. Pour chaque document du corpus d'entrainement, nous disposons d'un titre, d'un résumé, ainsi que d'une liste de mots clefs attribués par un ingénieur documentaliste. Les mots clefs sont soit issus d'un thésaurus de spécialité, soit ajoutés par l'ingénieur documentaliste en fonction de leur pertinence.

L'édition 2012 présentait une tâche similaire, consistant à indexer des articles scientifiques par l'intermédiaire de mots clefs d'auteurs (Paroubek *et al.*, 2012). L'équipe ayant obtenu les meilleurs scores a choisi une approche mixte utilisant des espaces sémantiques (El Ghali *et al.*, 2012). L'approche que nous proposons pour le défi de cette année utilise également des espaces sémantiques. Néanmoins, nous avons pour notre part choisi d'aborder cette tâche de façon purement statistique, et, mis à part le passage en minuscule et la suppression des accents, nous n'utilisons aucun prétraitement. Cela nous permet d'avoir un protocole facilement applicable à des langues peu dotées.

Les représentations vectorielles que nous utilisons ont été générées par l'algorithme NCISC présenté dans la section suivante. Nous présenterons ensuite les deux approches que nous avons définies avant de discuter les résultats dans la section 4.

2 L'algorithme NCISC

NCISC [1] est un algorithme stochastique de factorisation de matrices de la famille de RandNLA (Drineas & Mahoney, 2016), performant sur de grandes dimensions. Il permet d'apprendre des représentations vectorielles sur des tâches non supervisées ou semi-supervisées. Ces vecteurs peuvent être utilisés directement dans des recherches de plus proches voisins ou bien servir d'étape de prétraitement pour d'autres tâches d'apprentissage.

A l'instar de la LSA (Landauer & Dumais, 1997), NCISC effectue une réduction de dimension en réalisant une factorisation d'une grande matrice creuse en trois matrices denses : une matrice contenant des représentations vectorielles pour la première dimension de la matrice, une matrice contenant les valeurs propres des dimensions latentes et une matrice contenant des représentations vectorielles pour la seconde dimension de la matrice. Cette réduction est effectuée de manière itérative. Des connaissances a priori peuvent être introduites au cours du processus sur chacune des dimensions sous forme de vecteurs pré-calculés qui sont concaténés aux matrices en cours de calcul. Cela permet par exemple d'infléchir le calcul afin que deux objets appartenant à une même catégorie (ou deux mots utilisés dans un même contexte) aient au final des représentations approchantes.

Pour ce défi, deux types de matrices ont été utilisés : une matrice mot-graphèmes pour construire des représentations vectorielles graphiques et une matrice document-mots pour construire des représentations vectorielles sémantiques.

3 Stratégies de sélection des mots clefs

3.1 Approche par modèle graphique

Nous avons émis une première hypothèse formulée ainsi : le titre et le résumé d'un article permettent de connaître la teneur de celui-ci. De cette hypothèse découle une règle de sélection des mots clefs à partir du thésaurus thématique correspondant : un mot clef est un terme du thésaurus dont tous les mots pleins qui le composent apparaissent dans le titre ou le résumé d'un article (recherche exacte). Les scores obtenus en recherchant les mots clefs de cette manière sont assez faibles, surtout pour

1. L'algorithme NCISC est développé par la société eXenSa dans son moteur eXenGine (http ://www.exensa.com/about-us/). Une démonstration d'eXenGine analysant les données de wikipédia anglais est disponible à l'adresse wikinsights.org.

le rappel. En effet, un mot ou un expression clef peut apparaître sous une forme différente dans le titre ou le résumé d'une part, et le thésaurus d'autre part. Or, étant donné que nous n'utilisons pas de prétraitement linguistique comme la lemmatisation, les mots au pluriel (ou accordés au féminin) par exemple sont donc considérés comme des mots différents d'une version au singulier (resp. accordée au masculin). Pour que les différentes versions d'un mot puissent être rapprochées, nous avons construit un modèle de représentation vectorielle graphique. Un mot est décomposé en graphèmes de 2 à 5 lettres ordonnées. Nous utilisons des marqueurs de début de mot mais pas de marqueur de fin, afin de ne pas donner de poids à la désinence d'un mot. Par exemple le mot « laches » est composé des graphèmes suivants :

```
[::s::l; la; ac; ch; he; es;
 ::s::la; lac; ach; che; hes;
 ::s::lac; lach; ache; ches;
 ::s::lach; lache; aches]
```

On compose ainsi une matrice mot-graphèmes dont on réduit la dimension à l'aide de NCISC afin d'obtenir des vecteurs de représentation graphique des mots. Ainsi, les vecteurs de « lexique » et « lexiques » ont une similarité très élevée entre eux, mais sont également proches du vecteur représentant le mot « lexical ».

Munis de ces vecteurs graphiques, nous pouvons définir une nouvelle règle de sélection des mots clefs ainsi : chaque terme du thésaurus dont tous les mots pleins qui le composent ont une similarité suffisante avec un terme du titre ou du résumé d'un article est sélectionné comme mot clef (recherche approchée).

	Précision	Rappel	F-mesure
Archéologie	42,12	44,09	41,12
Linguistique	20,36	24,62	21,29
Sciences de l'information	18,86	22,12	19,34
Chimie	23,85	15,36	17,15

TABLE 1 – Résultats obtenus par recherche approchée de mots clefs dans les titres et résumés en utilisant des représentations vectorielles graphiques et un seuil de similarité de sélection de $0,98$.

Le tableau 1 présente les scores obtenus avec une recherche approchée. Cette approche permet une forte augmentation du rappel ainsi qu'une faible baisse de la précision (en moyenne). L'amélioration en terme de F-mesure va de $9,39$ à $14,68$ par rapport à l'approche exacte.

3.2 Approche par modèle sémantique

L'approche précédente, par construction, ne permet d'attribuer que des mots clefs déjà présents dans le titre ou le résumé d'un document. A l'inverse, l'intérêt des espaces sémantiques est justement de prendre en compte des informations de similarité sémantique non explicites. Notre seconde approche utilise ce type de représentation : à partir d'un modèle sémantique issu d'un apprentissage, nous construisons une représentation vectorielle d'un document et nous l'utilisons pour trouver les mots clefs qui lui sont associés. L'approche suivie est la suivante : pour un document donné, nous déterminons quels sont les documents les plus similaires puis, connaissant leur mots clefs respectifs, nous en déduisons les mots clefs liés au document. Il serait possible de chercher directement les mots

clefs associés à un document, sans passer par la recherche de documents voisins, mais cette approche ne donne pas les meilleurs résultats. D'autres travaux ont déjà utilisé l'information contenue dans des documents proches pour l'extraction de mots clefs, comme (Wan & Xiao, 2008) qui réalisent une partition thématique avant de repérer les mots saillants à l'aide d'un algorithme de graphe.

Le modèle sémantique utilisé est une matrice document-mots représentant les documents des corpus en sacs de mots d'uni-, bi- et tri-grammes, sélectionnés selon un critère d'information mutuelle. La dimension de la matrice est réduite à l'aide de l'algorithme NCISC. Au cours de la réduction, nous introduisons deux a prioris selon la procédure présentée dans la section 2 :

— un a priori sur les documents, généré à partir d'une matrice document-mots clefs,
— un a priori sur les mots, généré à partir d'une matrice de cooccurrences issue de la base wikipédia en français.

3.2.1 Vecteur de réprésentation des documents tests

Après apprentissage, chaque document du corpus d'entraînement dispose d'une représentation vectorielle tout comme chaque mot apparaissant dans le corpus. Pour qu'il soit possible de comparer les documents du corpus de test aux documents du corpus d'apprentissage, chaque document du corpus de test doit être représenté par un vecteur compatible. La représentation d'un document est contruite à partir des vecteurs des mots qu'il contient, qui sont sommés membre à membre. Il faut ensuite ôter de ce vecteur les valeurs propres de la matrice document-mots réduite, calulées sur le corpus d'entraînement par NCISC, afin d'obtenir une représentation comparable aux vecteurs des documents d'entraînement. Pour estimer la similarité entre deux documents dans cet article, nous utilisons le complément à 1 de la distance cosinus entre leur représentation vectorielle.

Il est à noter que nous aurions pu apprendre les représentations des mots et des documents sur un corpus élargi constitué à la fois des documents d'entraînement et des documents de test. Entraînées sur plus de données, les représentations vectorielles des tests auraient sans doute été plus précises. Nous avons cependant choisi de rester dans un cas d'utilisation où, quand de nouveaux documents se présentent, il est possible de trouver leurs mots clefs sans avoir besoin de refaire l'apprentissage.

3.2.2 Attribution des mots clefs via les plus proches voisins

Après le calcul de la représentation de chaque document du corpus de test, l'attribution des mots clefs s'effectue de la manière suivante :

— identification des K plus proches voisins parmi les documents du corpus d'entrainement de la spécialité associée,
— suppression des documents voisins ayant une similarité inférieure à un seuil S avec le document test,
— remplacement des documents voisins par leurs mots clefs associés, pondérés par leur similarité respective avec le document test. Si un même mot clef est lié à plusieurs documents voisins du document test, alors les scores de similarité sont additionnés,
— réordonnancement des mots clefs candidats en fonction de leur popularité dans le corpus d'entraînement : leur score de similarité est multiplié par le logarithme du nombre d'apparitions du mot clef dans le corpus d'entraînement.

Les scores présentés dans le tableau 2 montrent que cette approche favorise la précision au détriment du rappel. Les différents paramètres (dimension de la représentation vectorielle, nombre de documents

voisins sélectionnés avec seuil et nombre final de mots clefs proposés) ont été optimisés pour maximiser la F-mesure. La dernière colonne du tableau indique la précision au rang 1. La performance

	Précision	Rappel	F-mesure	Précision au rang 1
Archéologie	63,57	13,35	20,71	76,28
Linguistique	30,26	22,38	24,35	46,98
Sciences de l'information	40,69	11,62	15,88	42,45
Chimie	31,14	7,94	11,07	39,57

TABLE 2 – Résultats obtenus par similarité sémantique entre documents voisins

en terme de rappel est par contre bien moindre qu'avec le modèle précédent. C'est pourquoi nous avons cherché à combiner les deux solutions précédentes.

3.3 Combinaison des approches précédentes

Chacune des approches précédentes fournit une liste de mots clefs candidats associés à un score de similarité. Cependant, ces scores étant issus de deux modèles différents, calculés sur deux matrices différentes, ils ne sont pas directement comparables. Les scores de similarité des mots clefs issus du modèle graphique, plus élevés, sont donc multipliés par le score de similarité du meilleur mot clef issu du modèle sémantique. Les résultats sont ensuite fusionnés. Si un mot clef est proposé à la fois par le modèle sémantique et le modèle graphique, alors leurs scores sont additionnés. Une fois la liste finale ordonnée, les N premiers parmi ceux dont le score final dépasse un certain seuil de sélection sont conservés. Cette simple statégie de fusion donnant déjà de bons résultats, il serait intéressant par la suite d'en tester des plus élaborées.

Nous avons choisi de développer notre système avec peu de paramètres spécifiques au domaine. Cela permet une certaine stabilité pour de nouveaux domaines. De plus, notre approche statistique bénéficie d'un plus grand corpus d'apprentissage. Les modèles sont donc entraînés sur les quatre corpus d'entraînement réunis et les différents seuils ont été choisis pour maximiser la F-mesure moyenne sur les quatre corpus de développement. La dimension des vecteurs de représentation une fois les matrices réduites a ainsi été empiriquement fixée à 130. Seul le nombre maximal N de mots clefs finalement proposés dépend du type de corpus.

	Précision	Rappel	F-mesure	N
Archéologie	43,48	52,71	45,59	44
Linguistique	23,28	32,73	26,30	22
Sciences de l'information	21,26	30,32	23,86	29
Chimie	24,92	21,73	21,46	47

TABLE 3 – Résultats obtenus en combinant les deux approches avec un seuil de sélection de 0,4.

La précision obtenue avec ce modèle mixte est moindre que celle obtenue avec le modèle sémantique mais reste au dessus de celle du modèle graphique. Le véritable apport de cette combinaison est sur le score de rappel. En effet, bien que le modèle sémantique ait un rappel faible, son ajout augmente le rappel déjà élevé du modèle graphique. Le gain observé va de $6,37$ à $8,62$ points selon les cas. Cela confirme que l'utilisation de représentations vectorielles entraînées sur de grands corpus permet de capter de l'information sémantique qui n'est pas explicite dans les documents.

4 Discussion des résultats

Lors de ce défi, sur cinq participants, notre méthode s'est classée première sur les corpus de chimie et d'archéologie, deuxième sur les sciences de l'information et sur la linguistique, validant notre approche statistique. Nos représentations vectorielles sémantiques permettent bien de récupérer des mots clefs non présents dans les documents. Cependant, ce type d'approche de ne permet pas d'attribuer à un document un mot clef qui n'apparaît pas dans le corpus d'entraînement. La proportion de mots clefs du corpus d'apprentissage effectivement présents dans le corpus de test va donc avoir un impact sur les résultats. Cette proportion dépasse en effet en moyenne les 80 % dans les corpus d'archéologie et de sciences de l'information qui obtiennent de hauts scores de précision avec l'approche sémantique.

Des trois approches présentées dans cet article, l'approche sémantique est celle à privilégier pour obtenir une forte précision. A l'inverse, les approches graphiques et mixtes donnent un rappel meilleur que la précision pour trois des quatre corpus. Seul le corpus de chimie fait exception. L'hypothèse de l'approche graphique est que les mots clefs apparaîssent dans le titre ou le résumé. Les représentations vectorielles graphiques permettent d'intégrer certaines variations, en particulier, les graphèmes n'ont pas de marqueur de fin afin de donner peu de poids aux désinences. Par contre, la présence d'un marqueur de début rend difficile de rapprocher des mots qui diffèrent par leur préposition. Or en chimie, beaucoup de mots clefs sont des noms de molécules ou composés chimiques dont les dénominations se font par agglomération. Un texte peut contenir le mot « hexadécylsulfonate » et être indexé par le terme plus générique « sulfonate » qui ne sera donc pas reconnu. Il serait possible d'essayer un modèle entraîné sur des graphèmes ne comportant pas ce marqueur de début. Un autre problème survient lorsqu'un mot clef comprend un mot générique du domaine qui est absent du résumé. Par exemple, un texte peut être indexé par « cellule tumorale » et contenir « tumeurs » mais pas « cellule ». Ces mots, très fréquents dans un domaine, pourraient être repérés à l'aide d'un tf-idf et ne pas être recherchés dans les notices pour la validation des mots-clefs les contenant.

5 Conclusion

Cet article présente la contribution d'eXenSa à DEFT 2016 sur la tâche d'indexation de notices bibliographiques par des mots clefs. Notre approche statistique combine deux parties, l'une graphique et l'autre sémantique. La première cherche dans la notice du document des mots graphiquement proches des mots clefs utilisés dans les corpus d'entraînement. La seconde attribue à un nouveau document les mots clefs associés aux documents du corpus d'entraînement qui lui sont sémantiquement les plus proches. Les deux approches utilisent des représentations vectorielles apprises en utilisant notre algorithme NCISC, un algorithme stochastique de factorisation de matrices. À des fins de généralisation, nous utilisons le même corpus d'entraînement et le même paramétrage pour tous les domaines de spécialité. Notre système final se classe premier sur deux des quatre corpus de test, deuxième sur les autres. De plus, une fois l'apprentissage effectué, l'attribution de mots clefs à un nouveau document ne prend que quelques millisecondes. Une approche statistique est donc bien apropriée pour cette tâche bien qu'elle soit ici pénalisée par la petite taille des corpus. Par construction, notre système n'attribue que des mots clefs déjà présents dans les corpus d'entraînement. La proportion de mots clefs du corpus d'apprentissage effectivement présents dans le corpus de test a donc un impact sur les résultats. Les conventions de mots clefs propres à chaque spécialité ont également une influence, notamment pour la partie graphique de notre méthode.

Références

DRINEAS P. & MAHONEY M. W. (2016). Randnla : randomized numerical linear algebra. *Communications of the ACM*, **59**(6), 80–90.

EL GHALI A., HROMADA D. & EL GHALI K. (2012). Enrichir et raisonner sur des espaces sémantiques pour l'attribution de mots-clés. volume 2012, p.77̃.

LANDAUER T. K. & DUMAIS S. T. (1997). A solution to plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, **104**(2), 211.

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.

WAN X. & XIAO J. (2008). 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 : Association for Computational Linguistics.

TopicRank en domaines de spécialité : participation du LINA à DEFT 2016

Adrien Bougouin Florian Boudin Béatrice Daille
LINA – UMR CNRS 6241, 2 rue de la Houssinière, 44322 Nantes Cedex 3, France
<prenom.nom>@univ-nantes.fr

RÉSUMÉ

Cet article présente la participation de l'équipe TALN du LINA au défi fouille de textes (DEFT) 2016. Développé pour l'indexation de mots-clés, notre système reprend une méthode à base de graphe état de l'art dans le domaine (TopicRank) et l'étend afin de mieux répondre aux attentes d'une indexation professionnelle réalisée dans le cadre d'une bibliothèque scientifique numérique. Notre système s'est classé à la troisième place sur un total de cinq participants.

ABSTRACT

LINA at DEFT 2016

This article presents the participation of the TALN group at LINA to the défi fouille de textes (DEFT) 2016. Developed specifically for automatic keyphrase annotation, our system improves an existing method (TopicRank), mimicking professional indexers of Digital Libraries. Our system ranked third out of a total of five systems.

MOTS-CLÉS : DEFT 2016, extraction de mots-clés, assignement de mots-clés, méthode à base de graphe, domaine de spécialité.

KEYWORDS: DEFT 2016, keyphrase extraction, keyphrase assignment, graph-based method, specific domain.

1 Introduction

L'indexation automatique consiste à identifier un ensemble de mots-clés (e.g. mots, termes) 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. L'édition 2016 du défi fouille de textes (DEFT) porte sur l'extraction automatique de mots-clés à partir d'articles scientifiques en français.

L'objectif du défi DEFT 2016 est de retrouver, à partir du contenu d'articles scientifiques, les mots-clés qui ont été attribués par des indexeurs professionnels. Les documents sont issus de quatre domaines de spécialité : la linguistique, les sciences de l'information, l'archéologie et la chimie. Cet article décrit le système que nous avons mis au point pour le défi.

Le reste de cet article est organisé comme suit. Nous commençons par présenter les données (cf Section 2), puis notre approche (cf Section 3) et enfin, nos résultats comparés à ceux des autres participants de DEFT 2016 (cf Section 4).

2 Données du défi DEFT 2016

Les données mises à disposition par les organisateurs du défi DEFT 2016 sont composées de quatre corpus traitant chacun d'un domaine de spécialité parmi la linguistique, les sciences de l'information, l'archéologie et la chimie. Chaque corpus consiste en un ensemble de notices bibliographiques (titre, résumé) aux formats TEI et TXT (texte pré-traité de la notice), et un thésaurus au format SKOS. Les corpus sont divisés en trois sous-ensembles : jeu d'apprentissage, de développement et de test. Pour les deux premiers jeux, nous disposons des mots-clés de référence assignés par des indexeurs professionnel de l'Inist.

3 Approche

Nous proposons une approche fondée sur celle de TopicRank (Bougouin & Boudin, 2014). TopicRank est une méthode à base de graphe pour l'extraction de mots-clés. Elle sélectionne d'abord des mots-clés candidats au sein du document à analyser, les groupes en sujets, projette les sujets dans un graphe et les ordonnent à la manière de TextRank (Mihalcea & Tarau, 2004). Contrairement à d'autres méthodes d'extraction de mots-clés, TopicRank est capable de réduire considérablement la redondance des mots-clés extraits.

Dans ce travail, nous reprenons TopicRank et en améliorons ses performances en tirant partie des éléments du domaine des collections de DEFT 2016. Notre approche utilise les mots-clés des notices d'entrainement comme éléments du domaine et n'utilise pas les thésaurus fournis par les organisateurs. De cette manière, son usage ne se restreint pas uniquement aux données similaires à celles présentées dans le cadre de DEFT 2016.

3.1 TopicRank

TopicRank repose sur cinq grandes étapes :

1. **Sélection des mots-clés candidats.** Suivant les travaux précédents (Wan & Xiao, 2008; Hasan & Ng, 2010), TopicRank sélectionne les plus longues séquences de noms et d'adjectifs en tant que mots-clés candidats :

$$\text{mots_cles_candidat} = (NOM|ADJ)+ \tag{1}$$

2. **Groupement en sujets.** TopicRank groupe les mots-clés candidats similaires en sujets. Deux candidats c_i et c_j sont jugés similaires lorsqu'ils partagent au moins un quart de leurs mots, racinisés d'après la méthode de (Porter, 1980) :

$$\text{sim}(c_i, c_j) = \frac{|\text{Porter}(c_i) \cap \text{Porter}(c_j)|}{|\text{Porter}(c_i) \cup \text{Porter}(c_j)|} \tag{2}$$

$$\forall c_i, c_j \in CANDIDATS, c_j \in \text{sujet}(c_i) \Rightarrow \text{sim}(c_i, c_j) \geq \frac{1}{4} \tag{3}$$

Le groupement est réalisé avec un algorithme de groupement hiérarchique agglomératif.

3. **Construction du graphe.** TopicRank représente le document par un graphe complet $G = (N, A \subseteq N \times N)$ où les nœuds N sont les sujets. Chaque sujet $n \in N$ est connecté aux autres par une arête pondérée $a \in A$ selon la force du lien sémantique entre les sujets :

$$\text{poids}(n_i, n_j) = \sum_{c_i \in n_i} \sum_{c_j \in n_j} \text{distance}(c_i, c_j) \tag{4}$$

$$\text{distance}(c_i, c_j) = \sum_{p_i \in \text{positions}(c_i)} \sum_{p_j \in \text{positions}(c_j)} \frac{1}{|p_i - p_j|} \tag{5}$$

   Plus faible est la distance entre les mots-clés candidats de deux sujets dans le document, plus élevé est le poids de l'arête entre les deux sujets.

4. **Ordonnancement des sujets.** À la manière de TextRank (Mihalcea & Tarau, 2004), TopicRank ordonne les sujets par importance selon le principe de recommandation. Plus un sujet est fortement connecté à un grand nombre de sujets, plus il gagne d'importance, et plus les sujets avec lesquels il est fortement connecté sont importants, plus l'importance qu'il gagne est forte :

$$\text{importance}(n_i) = (1 - \lambda) + \lambda \times \sum_{n_j \in A(n_i)} \frac{\text{poids}(n_i, n_j) \times \text{importance}(n_j)}{\sum_{n_k \in A(n_j)} \text{poids}(n_j, n_k)} \tag{6}$$

   Où λ est un facteur de lissage fixé à 0,85 par Brin & Page (1998).

5. **Extraction des mots-clés.** TopicRank extrait un unique mot-clé pour chacun des k plus importants sujets. Bougouin *et al.* (2013) ont choisi de sélectionner dans chaque sujet le mot-clé candidat qui apparaît en premier dans le document.

Notre approche modifie les étapes de construction du graphe, d'ordonnancement par importance et de sélection des mots-clés de TopicRank. La construction du graphe étend le graphe de sujet en l'unifiant à un graphe des mots-clés de référence du domaine. L'ordonnancement est désormais conjoint entre les sujets du document et les mots-clés du domaine. Enfin, la sélection des mots-clés ajoute la possibilité de puiser dans le graphe du domaine. De cette manière, notre méthode est capable de réaliser simultanément deux catégories d'indexation par mots-clés :

- **Extraction de mots-clés** : les mots-clés sont sélectionnés parmi les unités textuelles du document (e.g. TopicRank) ;
- **Assignement de mots-clés** : les mots-clés ne sont pas restreints au contenu du document et doivent faire partie d'un vocabulaire contrôlé construit pour cette tâche.

3.2 Extraction et assignement (M1)

Afin de réaliser simultanément extraction et assignement de mots-clés, nous unifions deux graphes : l'un représentant le document (graphe de sujets) et l'autre les mots-clés de référence de son domaine (graphe du domaine). Le premier graphe sert à l'extraction de mots-clés. Le second, construit à partir des mots-clés de référence de documents d'apprentissage, sert à l'assignement. Ainsi, nous faisons l'hypothèse que les mots-clés de référence des documents d'apprentissage constituent la terminologie du domaine et nous les utilisons comme substituts au vocabulaire contrôlé usuel en assignement de

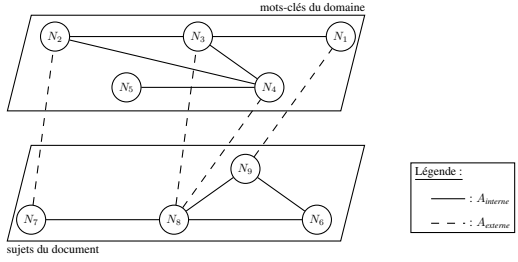

Figure 1: Illustration du graphe unifié que nous proposons

mots-clés. Contrairement aux mots-clés candidats sélectionnés dans le document, les mots-clés de référence ne sont pas redondants et ne sont donc pas groupés en sujets.

Soit le graphe unifié non orienté $G = (N, A = A_{interne} \cup A_{externe})$. N dénote indifféremment les sujets et les mots-clés du domaine. A regroupe les arêtes $A_{interne}$, qui connectent deux sujets ou deux mots-clés du domaine, et les arêtes $A_{externe}$, qui connectent un sujet à un mot-clé du domaine (voir la figure 1). Nous connectons deux sujets ou deux mots-clés du domaine lorsqu'ils apparaissent dans le même contexte et nous pondèrons leur arête par le nombre de fois que cela se produit. Lorsqu'il s'agit des sujets, le contexte est une phrase du document ; lorsqu'il s'agit des mots-clés du domaine, le contexte est l'ensemble des mots-clés du domaine d'un document d'apprentissage. Les contextes étant utilisés pour la création du graphe, le graphe de sujets n'est plus complet comme celui de TopicRank.

Le graphe de sujets et le graphe du domaine sont unifiés grâce aux arêtes $A_{externe}$. L'objectif des arêtes $A_{externe}$ est de connecter le document à son domaine par l'intermédiaire des concepts qu'ils partagent. Une arête $A_{externe}$ est donc créée entre un sujet et un mot-clé du domaine si ce dernier appartient au sujet, c'est-à-dire correspond à l'un de ses mots-clés candidats.

À partir du graphe unifié, nous ordonnons simultanément sujets $s \in N$ du document et mots-clés $m \in N$ du domaine par importance. Pour cela, nous reprenons le même principe que TopicRank et l'adaptons de sorte que sujets et mots-clés du domaine se transfèrent de l'importance. Nous proposons deux formes de recommandation : une recommandation interne $R_{interne}$ qui intervient entre deux nœuds du même type (sujets ou mots-clés du domaine) et une recommandation externe $R_{externe}$ qui intervient entre un sujet et un mot-clé du domaine.

$$\text{importance}(s_i) = (1 - \lambda_s)\ R_{externe}(s_i) + \lambda_s\ R_{interne}(s_i) \tag{7}$$

$$\text{importance}(m_i) = (1 - \lambda_m)\ R_{externe}(m_i) + \lambda_m\ R_{interne}(m_i) \tag{8}$$

$$R_{interne}(n_i) = \sum_{n_j \in A_{interne}(n_i)} \frac{\text{poids}(n_i, n_j)(n_j)}{\sum_{n_k \in A(n_j)} \text{poids}(n_j, n_k)} \tag{9}$$

$$R_{externe}(n_i) = \sum_{n_j \in A_{externe}(n_i)} \frac{\text{importance}(n_j)}{|A_{out}(n_j)|} \tag{10}$$

Où λ_s et λ_m sont deux facteurs de lissage définis empiriquement pour l'ordonnancement par importance des sujets et des mots-clés du domaine, respectivement.

Pour finir, indifféremment de leur type nous identifions dix mots-clés parmi les sujets et/ou mots-clés du domaine les plus importants.

3.3 Extraction seule (V1.1)

Nous proposons une variante de notre approche qui ne consiste qu'à identifier les mots-clés parmi les sujets. Le graphe unifié est toujours utilisé et les mots-clés du domaine sont tout de même utilisés pour l'ordonnancement.

3.4 Assignement seul (V1.2)

Nous proposons une seconde variante de notre approche qui ne consiste qu'à identifier les mots-clés parmi les mots-clés du domaine. Le graphe unifié est toujours utilisé et les sujets du document sont tout de même utilisés pour l'ordonnancement.

4 Résultats

Nous présentons dans cette section les résultats officiels de la campagne DEFT 2016. Nous avons soumis trois exécutions pour chaque collection : l'une pour notre approche (M1) et deux autres pour ses variantes (V1.1 et V1.2).

	Archéologie			**Chimie**			**Linguistique**			**Sciences de l'information**		
Méthode	Précision	Rappel	F-mesure	Précision	Rappel	F-mesure	Précision	Rappel	F-mesure	Précision	Rappel	F-mesure
M1	49,86	31,16	37,28	20,87	17,45	18,11	22,23	24,87	23,24	20,61	20,65	20,21
V1.1	43,63	26,63	32,17	15,77	13,10	13,60	13,77	15,56	14,47	15,67	15,87	15,39
V1.2	**53,77**	**33,46**	**40,11**	**21,15**	**17,54**	**18,28**	**23,16**	**25,85**	**24,19**	**21,93**	**21,83**	**21,45**

Table 1: Résultats de nos trois exécutions pour les collections d'archéologie, de chimie, de linguistique et de sciences de l'information

Le tableau 1 présente les résultats de nos trois exécutions pour les collections d'archéologie, de chimie, de linguistique et de sciences de l'information. Nous constatons que la variante V1.2 obtient les meilleurs résultats. Compte tenu de la nature des collections de données, cette observation semble normale. En effet, les notices sont indexées par des indexeurs professionnels utilisant principalement un vocabulaire controlé. Cela révèle toutefois que notre modèle ne fait pas suffisamment émerger les mots-clés du domaine pour M1.

Le tableau 2 présente, pour les collections d'archéologie, de chimie, de linguistique et de sciences de l'information, le classement des différentes équipes sur la base de la meilleure soumission. Notre soumission est classée au rang 3 sur 5.

5 Conclusion

Nous avons décrit la participation du LINA à DEFT 2016. Notre système est le résultat d'une analyse approfondie du mode de fonctionnement d'un indexeur professionnel de l'INIST. Il combine deux

	Archéologie		Chimie		Linguistique		Sciences de l'information	
Rang	Équipe	F-mesure	Équipe	F-mesure	Équipe	F-mesure	Équipe	F-mesure
1	EXENSA	45,59	EXENSA	21,46	EBSIUM	31,75	EBSIUM	28,98,
2	LIMSI	43,26	EBSIUM	21,07	EXENSA	26,30	EXENSA	23,86
3	**LINA**	**40,11**	**LINA**	**18,28**	**LINA**	**24,19**	**LINA**	**21,45**
4	EBSIUM	34,96	LIPN	15,31	LIPN	19,07	LIPN	15,34
5	LIPN	30,75	LIMSI	15,29	LIMSI	15,63	LIMSI	12,49

Table 2: Classement de DEFT 2016 sur la base des meilleures soumissions pour les collections d'archéologie, de chimie, de linguistique et de sciences de l'information. Notre classement est indiqué en gras.

graphes représentant le document à analyser et son domaine de spécialité, lui permettant d'extraire des mots-clés du document et d'en assigner à partir de son domaine. Ces derniers n'apparaissent pas nécessairement dans le document. Notre système s'est classé à la troisième place sur un total de cinq systèmes.

Parmi les trois variantes que nous avons proposées, celle qui ne réalise que l'assignement de mots-clés est la meilleure. Cela signifie que notre modèle hybride pour l'extraction et l'assignement simultanés de mots-clés n'est pas optimal. Nous envisageons d'étendre ce travail en affinant le schéma de connection des deux graphes afin de faciliter l'émergence des mots-clés à assigner.

Remerciements

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

Références

BOUGOUIN A. & BOUDIN F. (2014). TopicRank : ordonnancement de sujets pour l'extraction automatique de termes-clés. *TAL*, **55**(1), 45–69.

BOUGOUIN A., BOUDIN F. & DAILLE B. (2013). Topicrank: Graph-Based Topic Ranking for Keyphrase Extraction. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP)*, p. 543–551, Nagoya, Japan: Asian Federation of Natural Language Processing.

BRIN S. & PAGE L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, **30**(1), 107–117.

HASAN K. S. & NG V. (2010). Conundrums in Unsupervised Keyphrase Extraction: Making Sense of the State-of-the-Art. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING)*, p. 365–373, Stroudsburg, PA, USA: Association for Computational Linguistics.

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.

PORTER M. F. (1980). An Algorithm for Suffix Stripping. *Program: Electronic Library and Information Systems*, **14**(3), 130–137.

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.