Indexation libre et contrôlée d'articles scientifiques Présentation et résultats du défi fouille de textes DEFT2012

Patrick Paroubek[1] Pierre Zweigenbaum[1] Dominic Forest[2] Cyril Grouin[1]
(1) LIMSI–CNRS, Rue John von Neumann, 91403 Orsay, France
(2) EBSI, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal H3C 3J7, Canada
{pap,pz,grouin}@limsi.fr, dominic.forest@umontreal.ca

RÉSUMÉ

Dans cet article, nous présentons la campagne 2012 du défi fouille de texte (DEFT). Cette édition traite de l'indexation automatique par des mots-clés d'articles scientifiques au travers de deux pistes. La première fournit aux participants la terminologie des mots-clés employés dans les documents à indexer tandis que la seconde ne fournit pas cette terminologie, rendant la tâche plus complexe. Le corpus se compose d'articles scientifiques parus dans des revues de sciences humaines, indexés par leurs auteurs. Cette indexation sert de référence pour l'évaluation. Les résultats ont été évalués en termes de micro-mesures sur les rappel, précision et F-mesure calculés après lemmatisation de chaque mot-clé. Dans la piste fournissant la terminologie des mots-clés employés, la F-mesure moyenne est de 0,3575, la médiane de 0,3321 et l'écart-type de 0,2985 ; sur la seconde piste, en l'absence de terminologie, la F-mesure moyenne est de 0,2055, la médiane de 0,1901 et l'écart-type de 0,1516.

ABSTRACT

Controlled and free indexing of scientific papers
Presentation and results of the DEFT2012 text-mining challenge

In this paper, we present the 2012 edition of the DEFT text-mining challenge. This edition addresses the automatic, keyword-based indexing of scientific papers through two tracks. The first gives to the participants the terminology of keywords used to index the documents, while the second does not provide this terminology. The corpus is composed of scientific papers published in humanities journals, indexed by their authors. This indexing is used as a reference for the evaluation. The results have been evaluated in terms of micro-measures on the recall, precision and F-measure computed after keyword lemmatization. In the track giving the terminology of used keywords, the mean F-measure is 0.3575, the median is 0.3321 and the standard deviation is 0.2985 ; in the second track, the mean F-measure is 0.2055, the median is 0.1901 and the standard deviation is 0.1516.

MOTS-CLÉS : Campagne d'évaluation, fouille de textes, indexation libre, indexation contrôlée, mots-clés, thésaurus.

KEYWORDS: Evaluation campaign, Text-Mining, Free Indexing, Controlled Indexing, Keywords, Thesaurus.

JEP-TALN-RECITAL 2012, Atelier DEFT 2012: DÉfi Fouille de Textes, pages 1–13,
Grenoble, 4 au 8 juin 2012.

1 Introduction

La rédaction d'un article scientifique s'accompagne généralement de méta-données que l'auteur de l'article doit très souvent renseigner : titre, auteurs, affiliation des auteurs, et généralement un résumé présentant brièvement le contenu de l'article et un ensemble de mots-clés décrivant les thèmes de l'article. Ces mots-clés visent à aider la recherche des articles dans les bases de données bibliographiques.

La campagne DEFT 2012 s'intéresse à la détermination des mots-clés appropriés pour un article. Cela demande d'une part de déterminer les thèmes principaux de l'article et d'autre part de choisir des termes pour les nommer.

Certaines disciplines ont constitué un thésaurus qui prescrit les termes à employer pour cela. C'est le cas par exemple des sciences de la vie avec le thésaurus MeSH (Medical Subject Headings)[1] avec ses 26 142 descripteurs (version 2011), ou encore de l'informatique avec la classification hiérarchique de l'ACM[2] et ses 368 classes. Des thésaurus à vocation plus large ont aussi été créés, telle que la classification de la Bibliothèque du Congrès des États-Unis[3]. Ces référentiels visent à contrôler l'indexation des documents et à aider ainsi leur recherche.

À l'inverse, dans certaines revues ou conférences, aucun référentiel n'est imposé pour le choix des mots-clés indexant un article. Dans cette indexation libre, généralement réalisée par les auteurs eux-mêmes, le choix des mots-clés devient plus subjectif, chacun ayant une vision différente des termes à utiliser pour caractériser l'article.

C'est donc dans le cadre de cette problématique d'indexation libre ou contrôlée des articles scientifiques que nous avons inscrit cette nouvelle édition du défi fouille de texte (DEFT).

1.1 État de l'art

Les travaux dans le domaine de l'indexation des documents, qu'elle soit automatique (Salton *et al.*, 1975) ou non (Lancaster, 2003), ne sont pas récents. Parmi les méthodes généralement appliquées pour la création de classes de termes, celles-ci comprennent traditionnellement deux étapes : l'identification de termes dans un premier temps, puis la sélection des meilleurs candidats. L'étude des cooccurrences de termes (avec pré-traitements tels que étiquetage des parties du discours et lemmatisation) et l'utilisation de connaissances du domaine permet d'obtenir des résultats exploitables (Toussaint *et al.*, 1998). Ces techniques reprennent celles en vigueur en recherche d'information. Appliquées aux syntagmes nominaux, elles fournissent une base qui ne peut cependant suffire pour l'indexation (Sidhom, 2002). Des expériences d'indexation contrôlée automatique (au moyen de l'algorithme Okapi) et manuelle sur un corpus en français ont démontré l'intérêt de combiner ces deux approches pour améliorer les résultats (Savoy, 2005). Des approches plus récentes en matière d'indexation automatiques prennent en compte la coccurrence des termes associées à la structure des documents (Pompidor *et al.*, 2008). D'autres méthodes ont aussi été exploitées pour assister l'indexion automatique des documents, parmi lesquelles on retrouve la sémantique latente (Deerwester *et al.*, 1990).

[1] MeSH (National Library of Medicine) : `http://www.nlm.nih.gov/mesh/MBrowser.html`.

[2] Association for Computing Machinery, Computing Classification System : `http://dl.acm.org/ccs.cfm?part=author&coll=portal&dl=GUIDE`.

[3] Library of Congress Classification : `http://www.loc.gov/catdir/cpso/lcc.html`.

1.2 Déroulement

Un appel à participation a été lancé le 5 février 2012 sur les principales listes de diffusion dans les domaines des sciences de l'information *(ASIS-L)*, de la fouille de textes *(TextAnalytics, KDnuggets)*, des humanités numériques *(DH, Humanist)*, du Traitement Automatique des Langues et de la linguistique de corpus *(Corpora, LN, etc.)*. Dix-huit équipes se sont inscrites, pour certaines alors même que la phase de test avait déjà commencé, tandis que dix équipes ont poursuivi leurs efforts jusqu'à la période de tests. Ces équipes sont les suivantes, des inscriptions les plus anciennes (6 février) aux plus récentes (11 avril) :

- FBK, *Fondazione Bruno Kessler*, Trento, Italie : Sara Tonelli, Elena Cabrio, Emanuele Pianta.
- LIM&BIO, *Laboratoire d'Informatique Médicale & bioinformatique*, Université Paris 13 Nord, Bobigny (93) : Thierry Hamon.
- URPAH, *Unité de Recherche en Programmation Algorithmique et Heuristique*, Faculté des Sciences de Tunis, Tunisie : Amine Amri, Mbarek Maroua, Chedi Bechikh, Chiraz Latiri, Hatem Haddad.
- GREYC, *Groupe de Recherche en Informatique, Image, Automatique et Instrumentalisation de Caen*, Université de Caen Basse-Normandie, Caen (14) : Gaëlle Doualan, Mathieu Boucher, Romain Brixtel, Gaël Lejeune et Gaël Dias.
- IRISA, *Institut de Recherche en Informatique et Systèmes Aléatoires*, Université Rennes 1, Rennes (35) : Vincent Claveau et Christian Raymond.
- LINA, *Laboratoire d'Informatique de Nantes Atlantique*, Université de Nantes/École des Mines de Nantes, Nantes (44) : Florian Boudin, Amir Hazem, Nicolas Hernandez et Prajol Shrestha.
- LIMSI, *Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur*, Orsay (91) : Alexander Pak.
- LUTIN, *Laboratoire des Usages en Technologies d'Information Numérique*, Université Paris 8/UPMC/UTC/Universcience, Paris (75) : Adil El Ghali, Daniel Hromada et Kaoutar El Ghali.
- LORIA, *Laboratoire Lorrain de Recherche en Informatique et ses Applications*, Nancy (54) : Alain Lelu et Martine Cadot.
- PRISM, laboratoire *Parallèlisme, Réseaux, Systèmes et Modélisation*, Université Versailles–Saint-Quentin-en-Yvelines (78) et LaISC *Laboratoire d'Informatique et des Systèmes Complexes*, EPHE, Paris (75) : Murat Ahat, Coralie Petermann, Yann Vigile Hoareau, Soufian Ben Amor et Marc Bui.

Les corpus d'entraînement ont été diffusés aux participants inscrits ayant retournés l'accord de restriction d'usage des corpus signés à partir du 6 février 2012. Chaque équipe a choisi une fenêtre de trois jours durant la semaine du 9 au 15 avril 2012 pour appliquer ses méthodes sur le corpus de test. Les résultats ont été communiqués aux participants le 17 avril. La version finale des articles présentant les méthodes utilisées était attendue pour le 1er mai, pour un atelier de clôture le 8 juin 2012 pendant la conférence jointe JEP/TALN à Grenoble.

Pour la première fois dans l'histoire de DEFT, nous avons voulu mettre en place une interface de soumission des fichiers de résultats qui permettent de lancer une évaluation. Cette interface, dérivée d'une version utilisée dans un projet d'annotation de corpus, a nécessité de nombreuses adaptations et n'a pu être utilisée par les participants que trop tardivement (à partir du 6 avril, soit une semaine avant le démarrage de la phase de test) avec une fonction d'évaluation réellement opérationnelle qu'en fin de période de test. En conséquence, les participants au défi n'ont pas pu accéder à l'outil d'évaluation de leurs résultats pendant la période d'entraînement, ce qui, nous en convenons, ne facilite pas le développement de méthodes ni l'appréciation des évolutions offertes par les tentatives de modifications de ces méthodes durant cette période.

2 Présentation

Dans la continuité de l'édition 2011 du défi (voir DEFT2011), nous proposons de travailler de nouveau sur un corpus d'articles scientifiques parus dans le domaine des Sciences Humaines et Sociales. Alors que l'édition 2011 visait l'appariement de résumé avec l'article scientifique correspondant, nous proposons cette année d'identifier les mots-clés, tels qu'ils ont été choisis par les auteurs, pour indexer ces mêmes types d'articles. Les méthodes qui seront utilisées pour identifier les mots-clés devraient permettre de mettre en évidence les éléments saillants qui permettent d'indexer le contenu d'un article au moyen de mots-clés.

2.1 Pistes

Deux pistes sont proposées autour de l'identification de mots-clés (chaque piste dispose de ses propres corpus d'apprentissage et de test) :

– la première piste renvoie à l'indexation contrôlée des articles scientifiques et fournit la terminologie des mots-clés utilisés dans le corpus de cette piste (avec cependant une terminologie distincte pour chaque sous-corpus : une première pour l'apprentissage, une seconde pour le test), cette terminologie constituant une aide à la découverte des mots-clés ;
– la seconde piste renvoie à une indexation libre et ne fournit donc pas cette terminologie de référence ; les participants doivent identifier par eux-mêmes, dans le contenu du résumé et du corps de l'article, quels sont les mots-clés qui ont pu être choisis par l'auteur de l'article.

Sur chacune des deux pistes, le nombre de mots-clés indexant chaque document dans la référence est renseigné, tant dans le corpus d'apprentissage que dans le corpus de test. Les participants peuvent ainsi fournir exactement le nombre de mots-clés attendus.

Le travail d'indexation, qu'il s'effectue dans un cadre contrôlé ou non, reste complexe (Moens, 2000). Dans le cadre d'une indexation contrôlée, le choix de mots-clés parmi ceux proposés dans une terminologie reste difficile, l'indexeur, qu'il soit humain ou automatique, doit choisir parmi les termes proposés et uniquement parmi ceux-ci, les meilleurs candidats. Le travail consiste donc à identifier, parmi les termes proposés, quels sont ceux qui se rapprochent le plus de ceux que l'on aurait naturellement eu tendance à choisir. Dans le cadre d'une indexation libre, la première difficulté consiste à déterminer quels sont les meilleurs candidats à l'indexation, généralement en usant de méthodes statistiques éventuellement complétées par d'autres approches. Dans le cadre de ce défi, l'évaluation des termes qui auront été automatiquement choisis constitue une deuxième difficulté puisque la référence est constituée des mots-clés choisis par les auteurs des articles, ce choix étant purement subjectif mais considéré comme le meilleur pour cette campagne d'évaluation. Les résultats des participants sont donc évalués en comparaison d'une référence qui reste hautement perfectible.

Les participants peuvent participer, à leur convenance, aux pistes qu'ils souhaitent (seulement l'une ou les deux). Chaque participant est autorisé à soumettre jusqu'à trois fichiers de résultats par piste (soit un maximum de six exécutions pour une équipe participant aux deux tâches), permettant de tester officiellement trois systèmes ou trois configurations différentes d'un même système.

Les participants peuvent utiliser n'importe quelle ressource externe sauf celles provenant du site Erudit.org d'où proviennent les corpus.

2.2 Corpus

Le corpus se compose d'articles scientifiques provenant du portail Erudit.org parus entre 2003 et 2008 dans quatre revues de Sciences Humaines et Sociales : *Anthropologie et Société, Méta, Revue des Sciences de l'Éducation* et *Traduction, terminologie, rédaction*. Ces revues ont été sélectionnées car une majorité d'articles qui y ont paru sont accompagnés de mots-clés, choisis par les auteurs, indexant le contenu des articles. Ces mots-clés constituent la référence de cette édition, utilisée par les participants lors de la phase d'apprentissage et par les organisateurs pour évaluer les résultats lors de la phase de tests.

Du corpus initial de quatre revues, nous avons donc extrait 468 articles indexés par des mots-clés. Ces articles ont été répartis équitablement entre corpus des deux pistes, soit 234 articles par piste. Pour chaque piste, nous avons ensuite opéré une répartition entre corpus d'apprentissage et corpus de test selon le ratio 60/40% habituel, en nous assurant que ce ratio s'applique sur chaque revue (soit 60% des articles de chaque revue dans l'apprentissage et les 40% restants de chaque revue dans le test). Nous donnons ci-après (Figure 1) un exemple de document tel qu'il apparait dans le corpus d'apprentissage.

```
<doc id="0360">
 <motscles>
  <nombre>5</nombre>
  <mots>dimension;concept;caractère;spatiologie;organisation des connaissances</mots>
 </motscles>
 <article>
  <resume>
   <p>À partir de l'analyse de plusieurs termes se rapportant au domaine de la spatiologie, dans
      des langues aussi différentes que l'anglais et le français d'une part et l'arabe d'autre
      part, nous nous proposons de démontrer l'importance de la notion de pluridimensionnalité
      du concept dans l'organisation des connaissances et la classification des objets du monde.
       Ce faisant, nous aboutirons aussi à la conclusion que la structuration d'un domaine de
      spécialité, l'élaboration de son arborescence et surtout la formulation d'une définition
      dépendent principalement des caractères pris en compte dans l'appréhension des concepts,
      donc nécessairement de la « dimension » du concept.</p>
  </resume>
  <corps>
   <p>Nous savons que le concept est l'unité de base de toute analyse terminologique. Que celle-ci
       soit synchronique ou diachronique, portant sur le terme ou sur la définition, il faut
      toujours revenir au concept, à sa description au sein du système de concepts qu'il
      constitue avec les autres concepts appartenant au même domaine.</p>
   <p>Le concept est une « unité de connaissance créée par une combinaison unique de caractères »
      (ISO 1087-1 2000 : 2). Cette définition que donne la norme ISO 1087-1 2000 du concept met
      surtout l'accent sur la décomposition du concept en caractères, une décomposition qui
      permet une meilleure compréhension du concept et donc une meilleure organisation du
      système de concepts auquel il appartient.</p>
...
</corps>
</article>
</doc>
```

FIG. 1 – Extrait du corpus d'apprentissage avec méta-données associées

Chaque document intègre les éléments suivants :

– Des méta-données : la liste des mots-clés indexant le contenu de l'article (chaque mot clé est séparé du suivant par un point-virgule, information uniquement fournie dans les corpus d'apprentissage), mots-clés qu'il faudra identifier pour la phase de test (ligne 4) et le nombre de mots-clés indexant le contenu de l'article (information fournie dans les corpus d'apprentissage

et de test, ligne 3) ;
– L'article scientifique : le résumé de l'article (ligne 8) et le corps de l'article au complet (à partir de la ligne 11).

2.3 Terminologie

Sur la première piste, la terminologie des mots-clés employés dans le corpus est fournie (Figure 2). La terminologie du corpus d'apprentissage a été constituée en relevant tous les mots-clés des documents de ce corpus, classés par ordre alphabétique. La même procédure a été suivie pour constuire la terminologie du corpus de test. Puisque les mots-clés ont été choisis par les auteurs eux-mêmes, on constate que les mots-clés sont de différents types : des mots simples *(ethnologie)*, des mots composés *(Amérique latine)*, des expressions complexes *(Amérindien du Nord-Est)* et des combinaisons d'informations présentes dans l'article rassemblées sous un même « mot-clé » *(1982, droit constitutionnel canadien)*. Si la question de la difficulté de rattacher chaque mot-clé de cette terminologie aux documents du corpus se pose pour la première piste, les exemples présentés ici témoignent également de la difficulté à venir pour identifier les mots-clés sur la seconde piste, en l'absence de toute terminologie, compte-tenu de la grande variabilité des modalités de constitution des mots-clés.

```
1867, Constitution Act
1982, droit constitutionnel canadien
Abélès
Afrique
Afrique de l'Est
Agrawal
Algériens
Amazonie
Ambedkar
Amérindien du Nord−Est
Amérique latine
Ancien Régime
Aubrée
...
ethnicité
ethno−fiction
ethnographie
ethnographie multisites
ethnolinguistique
ethnologie
exogamie
...
```

FIG. 2 – Extrait de la terminologie du corpus d'apprentissage

3 Évaluation

Les mesures qui ont été retenues pour l'évaluation 2012 sont les mesures de précision, rappel, et F-mesure (Manning et Schütze, 1999), calculées avec une micro-moyenne (Nakache et Métais, 2005). Ce sont ces mesures qui ont été utilisées pour la piste 5 de la campagne SemEval-2010 : *Automatic Keyphrase Extraction from Scientific Articles* (Kim *et al.*, 2010).

Notons D l'ensemble des idenfiants de documents, K l'ensemble de tous les mots-clés utilisés par le système, W l'ensemble des mots-clés utilisés dans la base documentaire, les données hypothèse H (formule 1), c'est-à-dire l'ensemble des paires associant un identifiant de document à un mot clé fourni par le système participant et R les données référence (formule 2), c'est-à-dire l'ensemble des paires associant un identifiant de document à un mot clé issu de la base documentaire. Naturellement, pour un même identifiant de document, il peut exister plusieurs paires, aussi bien dans H que dans R, mais nous n'aurons pas de paire doublon au sein de l'un de ces ensembles, car les mots-clés seront alors différents. En effet, il n'y a aucun intérêt à annoter un document plusieurs fois avec le même mot-clé

$$H = \frac{(d, \text{Lem}(\text{Norm}(w))}{d \in D, w \in W, ((d,w1) \in H) \wedge ((d,w2) \in H)} \Rightarrow w1 \neq w2 \quad (1)$$

$$R = \frac{(a, \text{Lem}(\text{Norm}(k))}{a \in D, k \in K, ((a,k1) \in R) \wedge ((a,k2) \in R)} \Rightarrow k1 \neq k2 \quad (2)$$

Norm() est une fonction de normalisation de la typographie des mots-clé (normalisation de la casse) et Lem() est une fonction de lemmatisation des mots-clé.

L'ensemble des mots-clé correctement associés à un document par le système correspond au taux de vrais positifs (TP, formule 3), l'ensemble des mots-clé incorrectement associés à un document par le système correspond au taux de faux positifs (FP, formule 4) et l'ensemble des mots-clé non trouvés par le système correspond au taux de faux négatifs (FN, formule 5).

$$\text{TP} = H \cap R \quad (3) \qquad \text{FP} = \frac{H}{(H \cap R)} \quad (4) \qquad \text{FN} = \frac{R}{(H \cap R)} \quad (5)$$

La précision, le rappel et la F-mesure calculés en micro-moyenne correspondent aux formules 6 :

$$\text{Précision} = \frac{|H \cap R|}{|H|} \qquad \text{Rappel} = \frac{|H \cap R|}{|R|} \qquad \text{F-mesure} = \frac{(2 \times \text{p} \times \text{r})}{(\text{p} + \text{r})} \quad (6)$$

Notons que nous utilisons l'égalité stricte sur les mots-clés sans avoir recourt à une distance sémantique qui permettrait par exemple, de s'apercevoir que *recherche d'information* est plus proche de *fouille de données* que d'*algorithmique* afin de ne pas biaiser l'évaluation par rapport à une ontologie particulière. Nous avons également décidé de ne pas prendre en compte les recouvrements partiels de termes comme ayant une certaine validité pour éviter de recompenser un système qui retournerait *fouilles archélogiques* alors que la bonne réponse est *fouille de données*. Bien entendu, ce choix a pour résultat que la fourniture de l'hyponyme d'un terme au lieu du

terme sera considérée comme tout aussi fausse que la fourniture de n'importe quel autre terme. La production de mesures de performance complémentaires peut être envisagée à titre indicatif. Pour les résultats officiels de la campagne, seule la performance en F-mesure en micro-moyenne sera prise en compte.

4 Tests humains

Nous avons effectué des tests humains sur les deux pistes auprès des étudiants du parcours « *Ingénierie Multilingue* » du M2 Professionnel de l'INaLCO (formation sciences du langage avec une dominante traitement automatique des langues, étudiants d'origine étrangère avec pour certains une maîtrise moyenne de la langue française). Pour chaque piste, un sous-corpus composé de quatre fichiers chacun a été produit (un fichier issu de chacune des quatre revues utilisées dans le corpus global). Nous remercions chacun des étudiants pour le travail accompli.

4.1 Première piste, avec terminologie

Sur la première piste, puisque la terminologie des mots-clés employés dans les quatre articles composant le sous-corpus est disponible, une simple projection des mots-clés sur ce corpus au moyen d'une commande informatique[4] permet d'identifier dans quel fichier apparaît 14 des mots-clés de la terminologie. Sur ces 14 mots-clés, un seul est attribué à deux fichiers ; l'attribution de ce mot-clé au fichier qui compte le plus d'occurrences de ce terme permet une indexation correcte. Pour les 4 mots-clés restants qui n'ont pu faire l'objet d'une projection (généralement des mots-clés composés : *traduction française et allemande, Éducation multiculturelle, éducation intellectuelle*), une recherche d'un des termes composant le mot-clé permet d'identifier correctement l'article auquel il doit être associé. Cette technique, sur un sous-corpus limité, permet d'identifier 100% des indexations (F-mesure de 1,000).

4.2 Seconde piste, sans terminologie

Sur la seconde piste, aucune terminologie des mots-clés n'ayant été fournie, la tâche a été jugée plus complexe par les étudiants comme en témoignent les résultats obtenus (voir Tableau 1, F-mesure moyenne de 0,216 et médiane de 0,208). Afin de dresser grossièrement le contenu

	AM	BM	IP	LM	LT	NS	SK
Précision	0.250	0.200	0.167	0.118	0.292	0.292	0.208
Rappel	0.250	0.208	0.167	0.083	0.292	0.292	0.208
F-mesure	0.250	0.204	0.167	0.098	0.292	0.292	0.208

TAB. 1 – Évaluation des tests humains sur la seconde piste

de chaque article, un script qui extrait les tokens et les trigrammes de tokens utilisés dans le

[4] `grep -of termino_appr.txt piste1/testSans/* | sort | uniq`

document classés par fréquence d'utilisation décroissante a été mis à contribution. À charge pour les étudiants de s'inspirer de ces listes et de les confronter au contenu réel de l'article pour créer des mots-clés potentiels.

En conclusion, la seconde piste (sans terminologie) a été jugée difficile. Les mots-clés employés ne se retrouvent pas forcément à l'identique (*traduction française et allemande*) mais peuvent correspondre à une concaténation de plusieurs expressions (*traduction allemande* et *traduction française*). Il apparait par ailleurs que les mots-clés employés peuvent ne pas apparaitre dans le texte mais résulter d'une inférence (*Colombie Britannique* alors que le texte ne mentionne pas le nom de cette province mais celui d'une ville de cette province). Enfin, la redondance d'une thématique d'un même champ sémantique exprimée au moyen de deux mots-clés (*interprète* et *interprétation*) a été jugée complexe parce que contre-intuitif (un annotateur humain ayant tendance à choisir soit l'un, soit l'autre).

5 Méthodes des participants

La plupart des participants a considéré la première piste (avec terminologie) comme une tâche de recherche d'information dans laquelle les mots-clés constituent la requête à traiter.

Pour la seconde piste (absence de terminologie), les participants ont utilisés des outils d'extraction de mots-clés après avoir supprimé les mots non significatifs puis des méthodes de réordonnancement des mots-clés candidats. Concernant le niveau de granularité sur lequel travailler, une équipe (n° 04) a tenté le niveau caractère et le niveau mot (Doualan *et al.*, 2012) tandis qu'une autre équipe (n° 03) a fait le pari de travailler uniquement à l'échelle du syntagme nominal, considérant qu'un terme complexe est moins ambigü qu'un terme simple isolé (Amri *et al.*, 2012).

Plusieurs outils d'extraction de termes ont ainsi été mobilisés : l'outil *KX* accorde ainsi un poids aux termes extraits selon des annotations linguistiques et des relevés statistiques (Tonelli *et al.*, 2012) (n° 01), l'outil *TermoStat* qui repose sur des méthodes symboliques puis effectue un tri statistique (Claveau et Raymond, 2012) (n° 05), l'algorithme *KEA* (Keyphrase Extraction Algorithm) utilisé par l'équipe 06 (Boudin *et al.*, 2012). Une équipe (n° 02) a utilisé des outils de constitution de terminologies structurées pour reconnaître les termes (librairie *TermTagger* en Perl) extraire les termes (outil *YaTeA*) (Hamon, 2012). Les participants ont généralement utilisé des méthodes de pondération des mots-clés extraits reposant principalement sur le tf*idf, parfois en complétant avec la position du mot dans le document (Boudin *et al.*, 2012; Claveau et Raymond, 2012; Doualan *et al.*, 2012; Hamon, 2012; Tonelli *et al.*, 2012), la fréquence dans l'article, dans le résumé, la longueur de la chaîne, la présence du terme dans l'introduction et la conclusion (Doualan *et al.*, 2012). Certaines équipes ont également travaillé sur la reconnaissance des variantes morpho-syntaxiques des termes candidats (Hamon, 2012; Claveau et Raymond, 2012) en utilisant notamment l'outil *Fastr*. Mais l'approche qui a permis d'obtenir les meilleurs résultats (El Ghali *et al.*, 2012) repose sur une combinaison de plusieurs modules linguistiques d'ordre morphologique, sémantique et pragmatique.

En ce qui concerne le choix des meilleurs candidats, le cosinus a généralement été employé (Ahat *et al.*, 2012; Hamon, 2012), parfois en combinaison avec d'autres techniques telles que les graphes par l'équipe 18 (Ahat *et al.*, 2012) ou les réseaux bayésiens (El Ghali *et al.*, 2012). D'autres techniques fondées sur l'apprentissage ont également été mobilisées.

6 Résultats des participants

À l'image des tests humains, les participants ont obtenu de meilleurs résultats sur la première piste (où la terminologie des mots-clés employés était fournie) que sur la seconde (absence de terminologie). Nous renseignons dans le tableau 2 des résultats obtenus par les participants pour chacun des fichiers soumis dans chacune des deux pistes. Nous intégrons également une évaluation dite « hors compétition » pour les fichiers reçus après la fin de la période de test ; ces résultats ne sont pris en compte, ni dans le classement final, ni dans les statistiques globales (moyenne, médiane, écart-type).

Équipe	Run	Tâche 1			Tâche 2		
		Précision	Rappel	F-mesure	Précision	Rappel	F-mesure
01 – FBK	1	0,2682	0,2682	0,2682	0,1880	0,1880	0,1880
	2	0,2737	0,2737	**0,2737**	0,1452	0,1446	0,1449
	3	0,1978	0,1974	0,1976	0,1901	0,1901	**0,1901**
02 – LIM&BIO	1	0,3985	0,3985	**0,3985**	0,1798	0,1798	0,1798
	2	0,3333	0,3333	0,3333	0,1612	0,1612	0,1612
	3	0,2253	0,2253	0,2253	0,1921	0,1921	**0,1921**
03 – URPAH	1	0,0857	0,0857	**0,0857**	0,0785	0,0785	**0,0785**
	2	—	—	—	0,0785	0,0785	0,0785
04 – GREYC	1	0,0507	0,1769	0,0788	0,0469	0,1777	0,0742
	2	0,1082	0,1322	0,1190	0,1108	0,1488	**0,1270**
	3	0,4144	0,4730	**0,4417**	—	—	—
05 – IRISA	1	0,8017	0,7002	**0,7475**	0,2087	0,2087	**0,2087**
	2	0,7114	0,7114	0,7114	0,1704	0,1694	0,1699
	3	0,6760	0,6760	0,6760	—	—	—
06 – LINA	1	0,3812	0,4004	**0,3906**	0,1788	0,2128	0,1943
	2	0,3759	0,3948	0,3851	0,1949	0,2355	**0,2133**
	3	0,3343	0,4097	0,3682	0,1643	0,1880	0,1753
13 – LIMSI	1	0,1378	0,1378	**0,1378**	0,1632	0,1632	**0,1632**
16 – LUTIN	1	0,4618	0,4618	0,4618	0,2438	0,2438	0,2438
	2	0,9480	0,9497	**0,9488**	0,3471	0,3471	0,3471
	3	0,7486	0,7486	0,7486	0,5880	0,5868	**0,5874**
17 – LORIA	1	0,0522	0,2737	0,0877	0,0446	0,2562	0,0759
	2	0,0745	0,1955	**0,1079**	0,0603	0,1736	**0,0895**
	3	0,0401	0,3147	0,0711	0,0350	0,3017	0,0627
18 – PRISM	1	0,0428	0,0428	**0,0428**	—	—	—
	2	0,0242	0,0242	0,0242	—	—	—
Évaluations hors compétition							
HC 03 – URPAH	1	0,1695	0,1695	0,1695	0,1203	0,1198	0,1201
HC 15 – NOOPSIS	1	0,4587	0,2067	0,2850	0,0969	0,0909	0,0938

Tab. 2 – Résultats des participants pour chaque soumission sur les deux pistes

La correspondance entre numéro d'équipe et article présentant les méthodes s'établit comme suit : 01 – FBK (Tonelli *et al.*, 2012), 02 – LIM&Bio (Hamon, 2012), 03 – URPAH (Amri *et al.*, 2012), 04 – GREYC (Doualan *et al.*, 2012), 05 – IRISA (Claveau et Raymond, 2012), 06 – LINA (Boudin *et al.*, 2012), 16 – LUTIN (El Ghali *et al.*, 2012), et 18 – PRISM (Ahat *et al.*, 2012).

Sur la première piste, nous constatons des écarts extrèmement importants entre participants, avec des F-mesures qui varient de 0,0242 à 0,9488 ! On observe également des écarts élevés entre les différentes soumissions d'un même participant variant du simple au quadruple. Sur cette piste, si l'on se fonde sur les meilleures soumissions de chaque équipe, la F-mesure moyenne est de 0,3575, la médiane de 0,3321 et l'écart-type de 0,2985.

Sur la seconde piste, les écarts entre participants sont moindres, les F-mesures variant de 0,0627 à 0,5874. On observe également qu'un grand nombre de participants obtient, sur la meilleure soumission de son système, une F-mesure qui varie autour de 0,2. En se focalisant sur la meilleure soumission de chaque participant, la F-mesure moyenne est de 0,2055, la médiane de 0,1901 et l'écart-type de 0,1516.

Nous renseignons dans le tableau 3 du nombre de mots-clés intégrés dans chaque fichier de soumission. Sur la première piste, 443 mots-clés étaient attendus tandis que la seconde en attendait 391. Nombreux sont les participants qui ont fournis autant de mots-clés que le nombre attendu (ce nombre étant renseigné dans les méta-données de chaque document à traiter). Deux équipes ont fait le choix de retourner davantage de mots-clés que le nombre attendu.

Équipe	01 – FBK			02 – LIM&BIO			03 – URPAH		04 – GREYC			05 – IRISA		
Run	1	2	3	1	2	3	1	2	1	2	3	1	2	3
Tâche 1	443	443	442	443	443	443	443	—	1786	657	519	375	443	443
Tâche 2	391	390	391	391	391	391	391	391	1748	650	—	391	388	—

Équipe	06 – LINA			13 – LIMSI	16 – LUTIN			17 – LORIA			18 – PRISM	
Run	1	2	3	1	1	2	3	1	2	3	1	2
Tâche 1	470	470	564	443	443	444	443	2725	1315	4134	443	443
Tâche 2	483	492	461	391	391	391	391	2697	1302	4092	—	—

TAB. 3 – Nombre de mots-clés renseignés par fichier et par exécution sur chaque piste

Le GREYC (équipe 04) d'abord, avec environ quatre fois plus de mots-clés sur la première exécution, environ une fois et demie de plus sur la seconde soumission et à peine 1,17 fois de plus sur la troisième. Rapporté aux résultats obtenus, la troisième soumission — parce qu'elle correspond globalement au nombre attendu de mots-clés — obtient les meilleurs résultats. Le LORIA (équipe 17) enfin, avec environ six fois plus de mots-clés sur la première exécution, environ 3 fois plus sur la seconde et 9,33 fois plus sur la troisième soumission. À l'image du GREYC, la soumission dont le nombre de mots-clés se rapproche de celui attendu obtient les meilleurs résultats. Pour ces deux équipes, ces stratégies permettent d'obtenir un rappel meilleur que la précision mais les valeurs calculées restent faibles.

7 Conclusion

Les tâches d'indexation, bien que réalisées depuis de nombreuses années, ne constituent plus des pistes exploratoires. À ce titre, les résultats obtenus par les participants sur cette campagne témoignent des écarts importants entre équipes, selon que l'équipe dispose d'un système d'indexation ou bien part uniquement d'un système de base.

Les participants ont mieux réussi la première piste que la seconde, parce qu'elle fournissait la terminologie des mots-clés employés dans les documents du corpus à traiter. La F-mesure moyenne passe de 0,3575 sur la première piste à 0,2045 sur la seconde avec des écart-types variant de 0,2985 à 0,1522 de l'une à l'autre. On constate également des écarts élevés (jusqu'à 0,5388 d'écart de F-mesure), pour une même équipe, entre la meilleure soumission sur chaque piste.

Compte-tenu des modalités d'évaluation, les stratégies visant à fournir davantage de mots-clés que le nombre attendu (cette information ayant été fournie dans les corpus d'apprentissage et de test) ont permis d'accroître le rappel au détriment d'une précision très faible.

Remerciements

L'interface de soumission et d'évaluation des résultats a été développée par Pierre Albert dans le cadre du projet DoXa (financement CapDigital, convention DGE n° 08 2 93 0888). Nous remercions les organisateurs des conférences JEP/TALN pour l'organisation logistique de l'atelier et l'ATALA pour la mise à disposition d'une salle.

Nous remercions les étudiants du M2 Professionnel « Ingénierie Multilingue » 2011/2012 de l'INaLCO pour les tests humains qu'ils ont effectués, leur permettant ainsi de découvrir l'une des étapes essentielles lors de l'organisation d'une campagne d'évaluation : *Alexandra Moraru, Benjamin Marie, Irina Poltavchenko, Leidiana Martins, Lévana Thammavongsa, Nazim Saadi, Sofiane Kerroua*.

Références

Ahat, M., Petermann, C., Hoareau, Y. V., Ben Amor, S. et Bui, M. (2012). Algorithme automatique non supervisé pour le deft 2012. *In Actes de l'atelier de clôture du huitième défi fouille de texte (DEFT)*, pages 73–79.

Amri, A., Mbarek, M., Bechikh, C., Latiri, C. et Haddad, H. (2012). Indexation à base des syntagmes nominaux. *In Actes de l'atelier de clôture du huitième défi fouille de texte (DEFT)*, pages 37–43.

Boudin, F., Hazem, A., Hernandez, N. et Shrestha, P. (2012). Participation du lina à deft 2012. *In Actes de l'atelier de clôture du huitième défi fouille de texte (DEFT)*, pages 65–72.

Claveau, V. et Raymond, C. (2012). Participation de l'irisa à deft2012 : recherche d'information et apprentissage pour la génération de mots-clés. *In Actes de l'atelier de clôture du huitième défi fouille de texte (DEFT)*, pages 53–64.

DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K. et HARSHMAN, R. (1990). Indexing by latent semantic analysis. *In Journal of the American Society for Information Science*, volume 41, pages 391–407.

DOUALAN, G., BOUCHER, M., BRIXTEL, R., LEJEUNE, G. et DIAS, G. (2012). Détection de mots-clés par approches au grain caractère et au grain mot. *In Actes de l'atelier de clôture du huitième défi fouille de texte (DEFT)*, pages 45–52.

EL GHALI, A., HROMADA, D. et EL GHALI, K. (2012). Enrichir et raisonner sur des espaces sémantiques pour l'attribution de mots-clés. *In Actes de l'atelier de clôture du huitième défi fouille de texte (DEFT)*, pages 81–94.

HAMON, T. (2012). Acquisition terminologique pour identifier les mots-clés d'articles scientifiques. *In Actes de l'atelier de clôture du huitième défi fouille de texte (DEFT)*, pages 29–35.

KIM, S. N., MEDELYAN, O., KAN, M.-Y. et BALDWIN, T. (2010). Semeval-2010 task 5 : Automatic keyphrase extraction from scientific articles. *In Proc. of SemEval*, pages 21–26, Stroudsburg, PA. Association for Computational Linguistics.

LANCASTER, F. W. (2003). *Indexing and abstracting in theory and practice*. Facet, London.

MANNING, C. D. et SCHÜTZE, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massashusetts.

MOENS, M. F. (2000). *Indexing and abstracting of Document Texts*. Kluwer Academic Publishers.

NAKACHE, D. et MÉTAIS, E. (2005). Evaluation : nouvelle approche avec juges. *In INFORSID*, pages 555–570, Grenoble.

POMPIDOR, P., CARBONNEILL, B. et SALA, M. (2008). Indexation de co-occurrences guidée par la structure des documents et contrôlée par une ontologie et l'exploitation du corpus. *In INFORSID'08*, Fontainebleau, France. Lavoisier-Hermès.

SALTON, G., WONG, A. et YANG, C. S. (1975). A vector space model for automatic indexing. *In Communications of the ACM*, volume 18, pages 613–620.

SAVOY, J. (2005). Indexation manuelle et automatique : une évaluation comparative basée sur un corpus en langue française. *In Actes de Coria*, pages 9–23, Grenoble.

SIDHOM, S. (2002). *Plate-forme d'analyse morpho-syntaxique pour l'indexation automatique et la recherche d'information : de l'écrit vers la gestion des connaissances*. Thèse de doctorat, Université Claude Bernard – Lyon I.

TONELLI, S., CABRIO, E. et PIANTA, E. (2012). Key-concept extraction from french articles with kx. *In Actes de l'atelier de clôture du huitième défi fouille de texte (DEFT)*, pages 19–28.

TOUSSAINT, Y., NAMER, F., DAILLE, B., JACQUEMIN, C., ROYAUTÉ, J. et HATHOUT, N. (1998). Une approche linguistique et statistique pour l'analyse de l'information en corpus. *In* ZWEIGENBAUM, P., éditeur : *Actes de TALN 1998 (Traitement automatique des langues naturelles)*, pages 1–10, Paris. ATALA.