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.