Acquisition terminologique pour identifier les mots clés d'articles scientifiques

Thierry Hamon
LIM&BIO (EA3969)
Université Paris 13
93017 Bobigny Cedex
France
thierry.hamon@univ-paris13.fr

RÉSUMÉ

Le défi Fouille de texte 2012 (DEFT2012) a pour objectif d'identifier automatiquement des mots clés choisis par les auteurs d'articles scientifiques dans les Sciences Humaines et Sociales. Une liste de termes constituée des mots clés est fournie dans la tâche 1. Pour participer à ce défi, nous avons choisi d'exploiter des outils dédiés à la constitution de terminologies structurées. Les termes obtenus sont aussi été triés et filtrés à l'aide de leur position, de méthodes de pondération statistiques et de critères linguistiques. Plusieurs configurations de notre système ont été définies. Nous avons obtenu une F-mesure de 0,3985 dans la tâche 1 et de 0,1921 dans la tâche 2.

ABSTRACT

Terminological acquisition for identifying keywords of scientific articles

The challenge DEFT2012 aims at automatically identifying the keywords chosen by the authors of scientific articles in the Humanities. A keyword list is provided within the track 1. We propose to exploit terminological acquisition approaches. The extracted terms are also sorted and filtered according to their position in the documents, weighting measures and linguistic criteria. We defined several configurations of our system. Our best F-measure for the track 1 is 0.3985 while for the track 2, the best F-measure is 0.1921.

MOTS-CLÉS : Mots clés, extraction de termes, mesure de pondération, filtrage de termes.

KEYWORDS: Keywords, Term Recognition, Weighting Measure, Term Filtering.

1 Introduction

L'association de mots clés à un document, notamment à un article scientifique, est un aspect important de l'indexation documentaire. L'objectif du défi fouille de texte 2012 (DEFT2012) est d'identifier automatiquement les mots clés choisis par les auteurs d'articles scientifiques en Sciences Humaines et Sociales. Deux tâches sont proposées. Dans la première tâche, une liste de mots clés est fournie. Il s'agit donc de retrouver les mots clés les plus pertinents pour un document donné, parmi ceux fournis. Dans la deuxième tâche, aucune liste de mots clés n'est fournie. Il s'agit alors d'identifier les mots clés pouvant être associés à chaque document. Dans

JEP-TALN-RECITAL 2012, Atelier DEFT 2012: DÉfi Fouille de Textes, pages 25–31,
Grenoble, 4 au 8 juin 2012.

les deux cas, le nombre de mots clés attendus est connu.

Après une description, à la section 2, du matériel utilisé, nous présentons les différentes approches et paramètres utilisés à la section 3. Nous décrivons ensuite, à la section 4, les différentes expérimentations qui nous permis d'obtenir les résultats présentés à la section 5.

2 Matériel

Nous avons à notre disposition un corpus pour les phases d'entraînement et de test de chaque tâche. De plus, pour la tâche 1, une liste des mots clés associés à l'ensemble des documents du corpus est mise à disposition.

Corpus Le corpus est composé d'articles scientifiques parus entre 2001 et 2008 dans des revues de Sciences Humaines et Sociales :

– Revue des Sciences de l'Éducation (RSE),
– Traduction, Terminologie, Rédaction (TTR),
– Anthropologie et Sociétés (AS),
– Meta (journal des traducteurs – META).

Le corpus est décomposé en quatre sous-corpus, constituant les ensembles d'entraînement et de test pour les tâches 1 et 2. Pour chaque tâche, le corpus d'entraînement comporte environ 1 million de mots (environ 60% de la totalité du corpus pour une tâche donnée), tandis que les corpus de test sont constitués de 680 668 mots pour la tâche 1 et 639 267 mots pour la tâche 2 (voir tableau 1). Les corpus d'entraînement étaient disponibles pendant environ 2 mois, tandis que les corpus de test devaient être traités en 3 jours.

	Entraînement				Test			
Revue	Tâche 1		Tâche 2		Tâche 1		Tâche 2	
	mots	doc.	mots	doc.	mots	doc.	mots	doc.
RSE	143 314	19	160 387	20	112 203	13	91 371	13
TTR	95 731	13	96 083	13	65 056	9	65 382	9
AS	459 467	56	435 146	56	297 006	38	269 535	37
META	334 238	52	339 051	52	206 412	34	212 979	34
Total	1 032 750	140	1 030 667	94	680 677	141	639 267	93

TABLE 1 – Description des corpus d'entraînement et de test pour les tâches 1 et 2 (nombre de mots et de documents).

Liste des mots clés Pour la tâche 1, nous disposons également de la liste de mots clés du corpus. Ainsi, lors de l'entraînement, nous disposons de 66 mots clés, et lors de la phase de test, de 478 mots clés.

3 Méthode

Afin d'identifier les mots clés de chaque document, nous avons choisi d'utiliser dans un premier temps des approches d'extraction et de reconnaissance terminologiques (section 3.1). Les résultats de cette première étape sont ensuite triés avec des méthodes de pondération des termes (section 3.2). Enfin, une étape de filtrage et de sélection des termes triés permet d'identifier les mots clés potentiellement les plus pertinents pour chaque document (section 3.3)

3.1 Acquisition terminologique

Dans cette première étape, nous avons exploité des méthodes de reconnaissance ou d'extraction de termes pour identifier des mots clés. Afin d'étendre la couverture de l'acquisition de termes et d'améliorer l'étape de filtrage, nous avons également acquis des variantes morpho-syntaxiques des termes extraits.

Reconnaissance de termes (TermTagger). Les mots clés fournis lors de la tâche 1 ont été projetés sur l'ensemble des documents du corpus de travail. Cette projection a été élargie en prenant en compte les lemmes de mots du corpus. Pour réaliser la reconnaissance des mots clés, nous avons utilisé le module Perl `Alvis::TermTagger` [1].

Extraction des termes (YaTeA). Afin d'identifier les mots clés, nous avons choisi d'utiliser une méthode d'extraction de termes. Les mots clés pouvant être des mots ou des groupes nominaux, nous conservons aussi bien les termes complexes que leurs composants simples. Cette approche a été utilisée lors de la tâche 2 (où aucune liste de mots clés n'est fournie) mais aussi lors de la tâche 1 pour étendre l'identification des mots clés.

Pour réaliser l'extraction de termes sur les corpus, nous avons utilisé YaTeA [2] (Aubin et Hamon, 2006). Cet outil terminologique a pour objectif d'extraire d'un corpus des groupes nominaux qui peuvent être considérés comme de termes candidats. Il fournit leur analyse syntaxique sous forme d'une décomposition en tête et modifieur. L'extraction des termes est réalisée sur des critères linguistiques (patrons d'analyse simples utilisés de manière récursive et combinés à une désambiguisation endogène et la prise en compte de phénomène de variation morpho-syntaxique). Des mesures de pondération statiques sont également associées à chaque terme (fréquence, TF-IDF, etc.).

Acquisition de variantes morpho-syntaxiques (Faster). L'acquisition de variantes morpho-syntaxiques nous permet d'étendre la couverture des termes extraits par YaTeA, mais aussi d'améliorer le tri et le filtrage des termes extraits.

Nous avons utilisé Faster (Jacquemin, 1997) en mode indexation libre. Il nous est ainsi possible de reconnaître les variantes des termes extraits par YaTeA à travers trois types de variation morpho-syntaxique : la coordination de termes, l'insertion et la juxtaposition de modifieurs

1. `http://search.cpan.org/~thhamon/Alvis-TermTagger/`
2. `http://search.cpan.org/~thhamon/Lingua-YaTeA/`

et la permutation. Comme nous ne disposons pas de ressources dérivationnelles, les variantes dérivationelles n'ont pas pu être identifiées.

3.2 Méthodes de pondération

Afin d'identifier les mots clés parmi les termes extraits du corpus, ceux-ci doivent être triés par ordre de pertinence. Pour cela, nous avons utilisé plusieurs méthodes de pondération :

- la fréquence du terme dans le document (**tf**)
- le **TF-IDF** associé au terme (Salton et McGill, 1986)
- la position de la première occurrence du terme (**position**). Nous considérons ici que les termes situés au début du document ont un poids plus élevés que ceux situés à la fin. La position est le nombre de caractères depuis le début du document. Les termes sont alors triés dans l'ordre décroissant. Nous n'avons pas distingué le résumé du reste du document afin d'avoir les termes présents dans le résumé parmi les premières positions.
- le cosinus de la position de la première occurrence du terme (**positionCos**). Nous avons fait l'hypothèse que les termes présents au début (notamment dans les sections *résumé* et *introduction*) ou à la fin (notamment *conclusion*) du document sont les plus pertinents. On place ainsi la première occurrence de chaque terme sur le cercle trigonométrique en considérant que le début du document est l'angle 0 et la fin l'angle 2π. Nous avons calculé le cosinus de l'angle formé par la position.
- l'origine des termes (**termOrigin**). Les termes issus de la liste de mots clés sont prioritaire sur les termes extraits par YaTeA ou les variantes morpho-syntaxiques. Dans le cas des termes extraits automatiquement, la pondération peut tenir compte de l'application du filtre **filtrTermino** (voir section 3.3).

3.3 Filtrage et sélection des termes

Les termes peuvent être filtrés ou regroupés suivant différents critères linguistiques :

- Suppression des termes situés dans des phrases rédigés dans une langue autre que le français (**filtrLang**). Les revues étant issues des Sciences Humaines et Sociales, les articles contiennent des phrases exemples pouvant être écrits dans différentes langues. L'extracteur de termes identifie alors des termes qu'il n'est pas nécessaire de considérer comme des mots clés. Pour identifier la langue des phrases des articles, nous avons utilisé le module Perl `Lingua::Identify`[3] en utilisant les paramètres par défaut et ne visant à identifier que le français, l'anglais, l'allemand, l'espagnol et l'italien.
- Suppression des termes (modifieurs de termes complexes) étiquetés comme adjectifs. Les adjectifs correspondant à des mots clés sont conservés (**filtrAdj**) lors de la tâche 1. Par exemple, *espagnol* est étiqueté comme un adjectif mais peut correspondre à un nom et donc à un mot clé. Il est alors conservé.
- Prise en compte de l'inclusion lexicale : les termes en position tête d'un terme et ayant de rang plus élevé dans la liste triée sont supprimés (**filtrInclLex**). Par exemple, dans la liste triée (*Verbes supports*, *traduction*, *paraphrase*, *traduction automatique*), le terme *traduction* sera supprimé car celui-ci est inclus dans le terme *traduction automatique*.

3. `http://search.cpan.org/~ambs/Lingua-Identify/`

– Regroupement des termes en fonction de leur forme canonique (concaténation des lemmes des composants) et filtrage par la forme fléchie la plus fréquente (**filtrCan**).
– Sélection des termes contenant au moins un mot plein issu de la liste des mots clés (**filtrTermino**). Nous faisons l'hypothèse que si les mots clés sont composés de mots caractéristiques du domaine (en écartant les mots vides), il est possible de conserver les termes composés de ces mots. Nous avons également associé à chaque terme, un poids correspondant à la proportion de mots caractéristiques présents parmi les mots composants le terme.

Les filtres **filtrAdj** et **filtrTermino** sont appliqués avant le tri de la liste de termes par ordre de pertinence, tandis que les autres sont utilisés avec la liste des termes triés. La liste résultante de ce filtrage est ensuite réduite au nombre de mots clés attendus pour chaque document.

4 Expérimentations

Nous avons réalisé plusieurs expériences afin d'identifier les combinaisons de paramètres les plus adaptés pour l'identification des mots clés. L'ensemble des traitements a été réalisé dans la plate-forme Ogmios (Hamon et Nazarenko, 2008). Chaque corpus a été segmenté en mots et en phrases. Nous avons utilisé le TreeTagger (Schmid, 1997) pour l'étiquetage morpho-syntaxique et la lemmatisation des mots. En fonction des paramètres des expériences, nous avons utilisé au moins un des trois outils terminologiques (TermTagger, YaTeA, Faster) décrits à la section 3.1. Les listes de termes ont ensuite été triées et filtrées en combinant plusieurs paramètres.

A partir des résultats sur le corpus d'entraînement, nous avons défini 3 runs pour les tâches 1 et 2. Pour tous les runs, nous avons effectué un filtrage sur la langue (**filtrLang**) et les adjectifs (**filtrAdj**).

Tâche 1 (utilisation de la liste des mots clés)

– Run 1 : Les termes sont identifiés à l'aide du **TermTagger** en exploitant la liste des mots clés fournie par le défi. Suite aux résultats sur le corpus d'entraînement, nous avons choisi de différencier les méthodes de filtrage en fonction des sous-corpus. Ainsi, pour les sous-corpus **RSE**, **TTR** et **META**, nous avons exploité la position des termes pour trier la liste, alors que pour le sous-corpus **AS**, les termes sont triés en fonction du produit des poids **positionCos** et **tf**. Le filtre **filtrInclLex** est ensuite appliqué pour réduire la liste des termes et sélectionner les mots clés.
– Run 2 : Nous avons exploité **TermTagger** et YaTeA pour extraire les termes du corpus. Le filtre **filtrTermino** a été appliqué pour d'une part sélectionner les termes les plus pertinents, d'autre part pour associer le poids **termOrigin** aux termes extraits par YaTeA. La liste des termes a ensuite été triée en fonction de la méthode de pondération **termOrigin** et, lorsque les valeurs sont égales, en fonction du poids **tf**.
– Run 3 : dans ce run, nous avons également exploité **TermTagger** et YaTeA pour extraire les termes du corpus et le filtre **filtrTermino**. Nous avons choisi d'utiliser différents poids pour le tri de la liste des termes en fonction des sous-corpus. Ainsi, pour les sous-corpus RSE et TTR, nous avons exploité le **TF-IDF** pour trier les termes. Pour les sous-corpus META et AS, nous utilisé le produit des poids **positionCos** et **tf**.

Tâche 2

– Run 1 : L'extraction des termes a été réalisé à l'aide de YATEA. Nous avons ensuite exploité le **TF-IDF** pour trier la liste des termes. Celle-ci a ensuite été réduite à l'aide du filtre **filtrCan** (regroupement des termes possédant les mêmes formes canoniques).
– Run 2 : Nous avons utilisé YATEA pour extraire les termes du corpus et Faster. La liste des termes a été triée en fonction du produit des poids **positionCos** et **TF-IDF**, et réduite à l'aide du filtre **filtrCan**.
– Run 3 : Les termes ont été extraits à l'aide de YATEA. Nous avons utilisé le produit des poids **positionCos** et **tf** pour trier les termes. Les termes ont ensuite été sélectionné à l'aide du filtre **filtrCan**.

5 Résultats et discussion

Les résultats obtenus sur le corpus de test sont présentés dans le tableau 2. Le meilleur run de la tâche 1 obtient une F-mesure de 0,39. Il s'agit de projeter les mots clés et de les trier en fonction de leur position. Les expérimentations sur les corpus d'entraînement ont montré que les paramètres liés à la position (**position** et **positionCos**) ont une influence sur les résultats. Les résultats obtenus sur les deux autres runs semblent montrer que la fréquence des termes dégradent l'identification des mots clés. Enfin, sur le corpus d'entraînement, nous avons observé que seulement 72 % des mots clés projetés avec **TermTagger** étaient présents dans le corpus [4], et que l'utilisation de YATEA permet d'augmenter légèrement les résultats (+0,5 %). De même, YATEA permet d'identifier 55 % des mots clés.

En ce qui concerne la tâche 2, nous obtenons une F-mesure de 0,19 pour le meilleur run. Il s'agit de trier les termes extraits avec YATEA, en fonction du produit des poids **positionCos** et de la fréquence dans le document, les termes étant ensuite regroupés et filtrés en fonction de leur forme canonique.

Run	tâche 1			Tâche 2		
	Précision	Rappel	F-mesure	Précision	Rappel	F-mesure
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

TABLE 2 – Résultats sur le corpus de test

L'analyse des résultats montre que lorsque la liste des mots clés est fournie, la position de la première occurrence de termes est prépondérante lors des phases de tri et de sélection. Par ailleurs, l'utilisation d'un extracteur de termes permet de couvrir correctement les mots clés à identifier.Enfin, les mesures de pondération globales telles que le TF-IDF ne permettent pas d'obtenir des résultats satisfaisants (le constat est le même avec la CValue (Maynard et Ananiadou, 2000)).

4. Il s'agit d'un calcul du rappel légèrement différent de celui utilisé par les organisateurs du défi.

6 Conclusion

Nous avons exploité des approches dédiées à l'acquisition terminologique pour identifier des mots clés dans des corpus d'articles scientifiques des Sciences Humaines et Sociales. Les listes de termes extraites ont été triées et filtrées à l'aide de méthodes de pondération (position, fréquence au sein du document, TF-IDF, etc.) et de critères linguistiques. Les résultats obtenus montrent l'importance de la prise en compte de la position dans cette tâche quel que soit le cas de figure. En revanche, une méthode de pondération globale comme le TF-IDF ne semble pas être très utile dans ce contexte applicatif.

Les approches d'acquisition terminologique et notamment les extracteurs de termes, permettent d'obtenir une couverture relativement correcte, mais il est nécessaire de poursuivre les investigations sur les mesures statistiques permettant de trier au mieux les termes extraits. Nous envisageons par la suite d'utiliser l'algorithme de Page Rank (Page *et al.*, 1998) pour trier et filtrer les termes. De même la structure des documents pourraient être exploités beaucoup plus, notamment en prenant en compte la présence des termes dans le résumé ou les différentes sections du document (le titre des documents n'était malheureusement pas disponible lors du défi, mais il nous semble qu'il pourrait être important de le prendre en compte). Enfin, l'identification automatique des mots clés pourrait être conçue comme une tâche d'assistance aux rédacteurs des documents. Dans ce cas de figure, il serait intéressant de pouvoir évaluer l'apport des différentes approches en calculant une précision sur les n premiers termes ou un pourcentage de la liste.

Références

AUBIN, S. et HAMON, T. (2006). Improving term extraction with terminological resources. *In* SALAKOSKI, T., GINTER, F., PYYSALO, S. et PAHIKKALA, T., éditeurs : *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, numéro 4139 de LNAI, pages 380–387. Springer.

HAMON, T. et 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.

MAYNARD, D. et ANANIADOU, S. (2000). Identifying terms by their family and friends. *In Proceedings of COLING 2000*, pages 530–536, Saarbrucken, Germany.

PAGE, L., BRIN, S., MOTWANI, R. et WINOGRAD, T. (1998). The pagerank citation ranking : Bringing order to the web. Rapport technique, Stanford Digital Library Technologies Project.

SALTON, G. et MCGILL, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA.

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