HAL open science

Influence des domaines de spécialité dans l'extraction de termes-clés

Adrien Bougouin, Florian Boudin, Béatrice Daille

To cite this version:

Adrien Bougouin, Florian Boudin, Béatrice Daille. Influence des domaines de spécialité dans l'extraction de termes-clés. Traitement Automatique des Langues Naturelles (TALN), Jul 2014, Marseille, France. pp.13-24. hal-01021452

HAL Id: hal-01021452

https://hal.science/hal-01021452

Submitted on 15 Jul 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Influence des domaines de spécialité dans l'extraction de termes-clés

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. Les termes-clés sont les mots ou les expressions polylexicales qui représentent le contenu principal d'un document. Ils sont utiles pour diverses applications, telles que l'indexation automatique ou le résumé automatique, mais ne sont pas toujours disponibles. De ce fait, nous nous intéressons à l'extraction automatique de termes-clés et, plus particulièrement, à la difficulté de cette tâche lors du traitement de documents appartenant à certaines disciplines scientifiques. Au moyen de cinq corpus représentant cinq disciplines différentes (archéologie, linguistique, sciences de l'information, psychologie et chimie), nous déduisons une échelle de difficulté disciplinaire et analysons les facteurs qui influent sur cette difficulté.

Abstract. Keyphrases are single or multi-word expressions that represent the main content of a document. Keyphrases are useful in many applications such as document indexing or text summarization. However, most documents are not provided with keyphrases. To tackle this problem, researchers propose methods to automatically extract keyphrases from documents of various nature. In this paper, we focus on the difficulty of automatic keyphrase extraction in scientific papers from various areas. Using five corpora representing five areas (archaeology, linguistics, information sciences, psychology and chemistry), we observe the difficulty scale and analyze factors inducing a higher or a lower difficulty.

Mots-clés : Extraction de termes-clés, articles scientifiques, domaines de spécialité, méthodes non-supervisées.

Keywords: Keyphrase extraction, scientific papers, specific domain, unsupervised methods.

1 Introduction

Un terme-clé est un mot ou une expression polylexicale qui représente un concept important d'un document auquel il est associé. En pratique, plusieurs termes-clés représentant des concepts différents sont associés à un même document. Ils forment alors un ensemble à partir duquel il est possible de caractériser, synthétiser, le contenu du document. Du fait de cette capacité de synthèse, les termes-clés sont utilisés dans de nombreuses applications telles que le résumé automatique (D'Avanzo & Magnini, 2005), la classification de documents (Han *et al.*, 2007) ou l'indexation automatique (Medelyan & Witten, 2008). Cependant, tous les documents ne sont pas accompagnés de termes-clés et leur assignation manuelle est une tâche coûteuse. Pour pallier ce problème, de plus en plus de chercheurs s'intéressent à l'extraction automatique de termes-clés, en témoignent les récentes campagnes d'évaluation (Paroubek *et al.*, 2012; Kim *et al.*, 2010), ainsi que les nombreux travaux à ce sujet (Hasan & Ng, 2014).

L'extraction automatique de termes-clés consiste à extraire du contenu d'un document les unités textuelles les plus importantes, celles qui permettent de le résumer. Parmi les méthodes d'extraction automatique de termes-clés existantes, nous distinguons deux catégories : les méthodes supervisées et les méthodes non-supervisées. Dans le cadre supervisé, la tâche d'extraction de termes-clés est considérée comme une tâche de classification (Witten *et al.*, 1999) où il s'agit d'attribuer la classe « *terme-clé* » ou « *non terme-clé* » aux termes-clés candidats du document. Une collection de documents annotés en termes-clés est utilisée pour l'apprentissage d'un modèle de classification reposant sur divers traits tels que la fréquence du terme-clé candidat ou sa position dans le document. Dans le cadre non-supervisé, les méthodes attribuent un score d'importance aux candidats selon divers indicateurs tels que leur degré de spécificité (Spärck Jones, 1972) ou les relations de coocurrence que leurs mots entretiennent (Mihalcea & Tarau, 2004). En général, les méthodes supervisées sont plus performantes que les méthodes non-supervisées, mais leur besoin en données d'apprentissage annotées et leur dépendance vis-à-vis du domaine de ces données d'apprentissage poussent les chercheurs à s'intéresser aux méthodes non-supervisées.

Dans cet article, nous nous plaçons dans le contexte de l'extraction non-supervisée de termes-clés à partir de documents de

nature scientifique. Faisant l'hypothèse que certaines disciplines sont plus difficiles à traiter que d'autres, nous présentons diverses stratégies d'extraction de termes-clés puis comparons leurs différences de performance. Nous déterminons ensuite quels sont les facteurs qui influent sur la difficulté de la tâche d'extraction automatique de termes-clés. De la connaissance de ces facteurs peut émerger le besoin d'utiliser des ressources externes, telles que des thésaurus, souvent mises de côté dans les travaux portant sur l'extraction non-supervisée de termes-clés. Cela peut aussi permettre de détecter la difficulté en amont de l'extraction de termes-clés afin d'affiner le paramétrage de la méthode utilisée.

Le reste de cet article est organisé comme suit. Dans un premier, temps nous présentons les collections de données (section 2) et les méthodes d'extraction de termes-clés (section 3) que nous utilisons. Dans un second temps, nous appliquons ces méthodes à nos collections de données (section 4), puis nous discutons des différents facteurs observables (section 5) avant de conclure (section 6).

2 Collections de données

Pour ce travail, nous disposons de cinq corpus disciplinaire de notices bibliographiques fournies par l'Inist [1] dans le cadre du projet ANR Termith [2] : archéologie, linguistique, sciences de l'information, psychologie et chimie. Chaque notice contient le titre, le résumé et les termes-clés d'un document auquel elle est associée. Les termes-clés sont classés en deux catégories :

- les termes-clés d'auteurs, assignés librement par les auteurs pour caractériser leur production ;
- les termes-clés Inist (en français, en anglais ou en espagnol), assignés par des indexeurs professionnels selon des règles précises destinées à améliorer la recherche d'information et à homogénéiser l'indexation des notices :
  - les termes-clés doivent être du même niveau de spécificité que celui du document et peuvent parfois être accompagnés d'un terme-clé plus générique pour le restituer dans son contexte ;
  - les termes-clés doivent respecter, autant que possible, le langage de la discipline à laquelle appartient le document (termes-clés contrôlés) ;
  - pour tous les documents d'une même discipline, un même concept doit être représenté par le même terme-clé ;
  - les termes-clés d'un document doivent présenter tous les concepts qui y sont importants, même ceux qui sont implicites.

Nous utilisons les termes-clés français assignés par l'Inist.

Le corpus d'**archéologie** est composé de 718 notices. Celles-ci représentent des articles parus entre 2001 et 2012 dans 22 revues différentes (*Paléo*, *Le bulletin de la Société préhistorique française*, etc.).

Le corpus de **linguistique** est constitué de 716 notices d'articles parus entre 2000 à 2012 dans 12 revues différentes (*Linx – Revue des linguistes de l'Université Paris Ouest Nanterre La Défense*, *Travaux de linguistique*, etc.).

Le corpus de **sciences de l'information** contient 706 notices d'articles publiés entre 2001 et 2012 dans six revues différentes (*Documentaliste – Sciences de l'information*, *Document numérique*, etc.).

Le corpus de **psychologie** contient 720 notices d'articles parus entre 2001 et 2012 dans sept revues différentes (*Enfance*, *Revue internationale de psychologie et de gestion des comportements organisationnels*, etc.).

Le corpus de **chimie** est composé de 782 notices d'articles publiés entre 1983 et 2012 dans quatre revues (*Comptes Rendus de l'Académie des Sciences*, *Comptes Rendus Chimie*, etc.).

Le tableau 1 présente les caractéristiques des cinq collections de données dont nous disposons. Les notices sont de petite taille et sont rédigées différemment selon les disciplines (cf. figure 1). Les notices d'archéologie, par exemple, font l'objet d'un effort de présentation du contexte historique lié aux travaux présentés, tandis que les notices de chimie, principalement des comptes rendus d'expériences, décrivent sommairement (énumèrent) les expériences réalisées (noms des expériences, éléments chimiques impliqués, etc.). Les termes-clés associés aux documents varient en nombre (de 8,5 à 16,6) et en complexité. Par exemple, en archéologie, nous observons qu'un grand nombre de termes-clés sont des entités nommées principalement composées d'un seul mot (p. ex. « Paléolithique », « Europe », etc.), tandis qu'en chimie, nous observons un usage fréquent de notions centrales (dans le langage de chimie) nécessitant une spécialisation systématique (p. ex. « réaction topotactique », « réaction sonochimique », « réaction électrochimique », etc.). Nous remarquons aussi

1. Institut de l'Information Scientifique et Technique : `http://www.inist.fr`
2. TERMinologie et Indexation de Textes en sciences Humaines : `http://www.atilf.fr/ressources/termith/`

Statistique	Archéologie	Linguistique	Sciences de l'information	Psychologie	Chimie
Documents	718	715	706	720	782
Mots/doc.	219,1	156,7	119,7	185,7	105,2
Termes-clés/doc.	16,6	8,0	8,5	11,6	12,8
Mots/terme-clé	1,3	1,8	1,7	1,6	2,2
Diversité des termes-clés	25,5 %	23,0 %	25,0 %	17,4 %	40,6 %
Termes-clés contrôlés	79,8 %	86,9%	85,8 %	90,9%	83,0 %
Termes-clés non contrôlés	20,2 %	13,1%	14,2 %	9,1%	17,0 %
Termes-clés extractibles (Rappel max.)	62,9 %	38,8 %	32,4 %	27,1 %	23,7 %
↪ Termes-clés contrôlés extractibles	48,8 %	34,9 %	27,9 %	24,9 %	21,7 %
↪ Termes-clés non contrôlés extractibles	14,1 %	3,9 %	4,5 %	2,2 %	2,0 %

TABLE 1 – Caractéristiques des corpus disciplinaires. La diversité des termes-clés représente la proportion de termes-clés différents dans la discipline $\left(\frac{\text{nombre de termes-clés différents}}{\text{nombre total de termes-clés}}\right)$. Les termes-clés extractibles sont les termes-clés pouvant être extraits du contenu des documents. Conformément au processus d'évaluation standard pour les méthodes d'extraction automatique de termes-clés (cf. section 4.1), les variantes flexionnelles d'un terme-clé de référence sont jugées correctes (p. ex. « langues de spécialité » peut être extrait à la place de « langue de spécialité »).

une diversité variable selon les disciplines (de 23,0 % à 40,6 % de termes-clés différents). En chimie, la diversité plus importante que pour les autres disciplines, c'est-à-dire un nombre plus important de termes-clés différents parmis tous les termes-clés de référence, indique une difficulté a priori plus importante. Enfin, il est important de noter la faible proportion de termes-clés apparaissant dans les notices, à une flexion près — rappel maximum pouvant être obtenu. Par exemple, dans le corpus de chimie, uniquement trois termes-clés peuvent être extraits des notices parmi les 12,8 associés aux notices, en moyenne, en comparant les candidats à partir de la racine de leurs mots déterminées avec la méthode de Porter (1980). Ce dernier point concerne principalement les termes-clés contrôlés, qui peuvent être assignés à un document à partir de règles concernant les unités textuelles présentent dans le document. Ces règles, dites de déclenchement, sont définies manuellement par les indexeurs professionnels et ne sont pas disponibles pour ce travail.

3 Extraction automatique de termes-clés

L'extraction non-supervisée de termes-clés peut se décomposer en quatre étapes (cf. figure 2). Tout d'abord, les documents sont un à un enrichis linguistiquement (segmentés en phrases, segmentés en mots et étiquetés en parties du discours), des termes-clés candidats y sont ensuite sélectionnés, puis ordonnés par importance et enfin, les k plus importants sont sélectionnés en tant que termes-clés. Les étapes les plus importantes d'un système d'extraction automatique de termes-clés sont celles de sélection des candidats et d'ordonnancement de ceux-ci. Intuitivement, l'ordonnancement des candidats est le cœur du système, mais la performance de celui-ci est limitée par la qualité de l'ensemble de termes-clés candidats qui lui est fourni. Un ensemble de candidats est de bonne qualité lorsqu'il fournit un maximum de candidats présents dans l'ensemble des termes-clés de référence et lorsqu'il fournit peu de candidats non-pertinents, c'est-à-dire des candidats qui ne sont pas dans l'ensemble des termes-clés de référence et qui peuvent dégrader la performance du système d'extraction de termes-clés utilisé.

3.1 Préparation des données

Les documents des collections de données utilisées subissent tous les mêmes prétraitements. Ils sont tout d'abord segmentés en phrases, puis en mots et enfin étiquetés en parties du discours. Dans ce travail, la segmentation en phrase est effectuée par le *PunktSentenceTokenizer* disponible avec la librairie Python NLTK (Bird *et al.*, 2009, *Natural Language ToolKit*), la segmentation en mots est effectuée par l'outil Bonsai du Bonsai PCFG-LA parser[3] et l'étiquetage en parties du discours est réalisé par MElt (Denis & Sagot, 2009). Tous ces outils sont utilisés avec leurs paramètres par défaut.

3. http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html

Variabilité du gravettien de Kostienki (bassin moyen du Don) et des territoires associés [a] *Archéologie*

Dans la région de Kostienki-Borschevo, on observe l'expression, à ce jour, la plus orientale du modèle européen de l'évolution du Paléolithique supérieur. Elle est différente à la fois du modèle Sibérien et du modèle de l'Asie centrale. Comme ailleurs en Europe, le Gravettien apparaît à Kostienki vers 28 ka (Kostienki 8 /II/). Par la suite, entre 24-20 ka, les techno-complexes gravettiens sont représentés au moins par quatre faciès dont deux, ceux de Kostienki 21/III/ et Kostienki 4 /II/, ressemblent au Gravettien occidental et deux autres, Kostienki-Avdeevo et Kostienki 11/II/, sont des faciès propres à l'Europe de l'Est, sans analogie à l'Ouest.

Termes-clés de référence : Europe*, Kostienko, Borschevo, variation*, typologie*, industrie osseuse*, industrie lithique*, Europe centrale*, Avdeevo*, Paléolithique supérieur*, Gravettien*.

Termes techniques et marqueurs d'argumentation : pour débusquer l'argumentation cachée dans les articles de recherche [b] *Linguistique*

Les articles de recherche présentent les résultats d'une expérience qui modifie l'état de la connaissance dans le domaine concerné. Le lecteur néophyte a tendance à considérer qu'il s'agit d'une simple description et à passer à côté de l'argumentation au cours de laquelle le scientifique cherche à convaincre ses pairs de l'innovation et de l'originalité présentées dans l'article et du bien-fondé de sa démarche tout en respectant la tradition scientifique dans laquelle il s'insère. Ces propriétés spécifiques du discours scientifique peuvent s'avérer un obstacle supplémentaire à la compréhension, surtout lorsqu'il s'agit d'un article en langue étrangère. C'est pourquoi il peut être utile d'incorporer dans l'enseignement des langues de spécialité une sensibilisation aux marqueurs linguistiques (terminologiques et argumentatifs), qui permettent de dépister le développement de cette rhétorique. Les auteurs s'appuient sur deux articles dans le domaine de la microbiologie.

Termes-clés de référence : Langue scientifique*, argumentation*, rhétorique*, langue de spécialité*, enseignement des langues*, linguistique appliquée*, discours scientifique*, article de recherche.

Etude d'un condensat acide isocyanurique-urée-formaldéhyde [c] *Chimie*

La synthèse d'un condensat acide isocyanurique-urée-formaldéhyde utilisant la pyridine en tant que solvant a été effectuée par réaction sonochimique.

Termes-clés de référence : Réaction sonochimique*, hétérocycle azote*, cycle 6 chaînons*, ether*.

a. http://cat.inist.fr/?aModele=afficheN&cpsidt=20563716
b. http://cat.inist.fr/?aModele=afficheN&cpsidt=17395748
c. http://cat.inist.fr/?aModele=afficheN&cpsidt=6719275

FIGURE 1 – Exemples de notices Inist. Les termes-clés soulignés sont ceux qui occurrent dans le titre ou le résumé de la notice. Les termes-clés marqués d'une * font partie des termes-clés contrôlés.

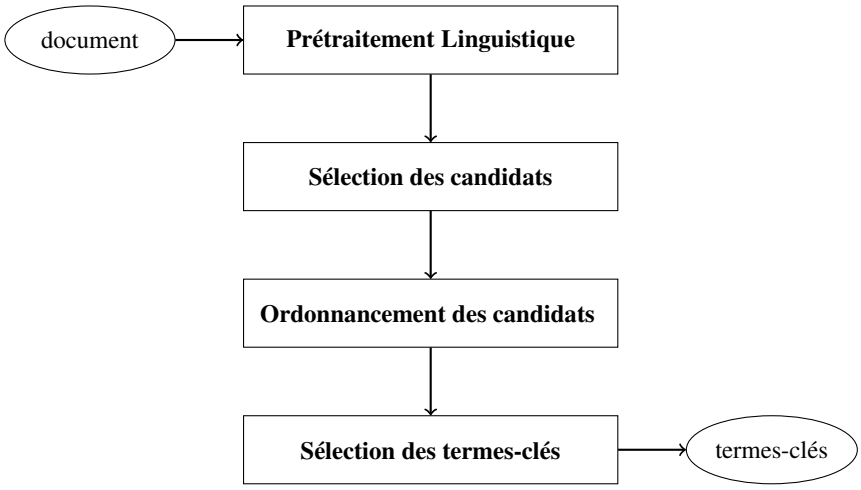


FIGURE 2 – Chaîne de traitements d'un système non-supervisé d'extraction automatique de termes-clés.

3.2 Sélection des termes-clés candidats

Dans les travaux précédents, deux approches sont fréquemment utilisées. Soit les méthodes sélectionnent les n-grammes (filtrés) en tant que termes-clés candidats, soit elles sélectionnent les candidats par reconnaissance de forme (Hulth, 2003). Dans ce travail, nous expérimentons trois méthodes différentes : deux méthodes conformes aux approches standards et une méthode sélectionnant les candidats termes obtenus par un extracteur terminologique. Aucun travail portant sur l'extraction automatique de termes-clés n'a, à notre connaissance, utilisé une telle approche. Compte tenu de la nature (disciplinaire) de nos données, nous faisons l'hypothèse que les candidats termes, tels que définis dans le domaine de l'extraction terminologique, peuvent aussi être des termes-clés candidats. Ces trois méthodes de sélection fournissent des ensembles de candidats de qualités différentes, ce qui nous permet par la suite d'identifier les facteurs qui influent sur la difficulté de l'extraction automatique de termes-clés.

La **sélection des n-grammes filtrés** consiste à extraire du document toutes les séquences ordonnées de n mots, puis à les filtrer avec un anti-dictionnaire regroupant les mots fonctionnels de la langue (conjonctions, prépositions, etc.) et les mots courants (« près », « beaucoup », etc.). Dans ce travail, nous suivons Witten *et al.* (1999) et sélectionnons les n-grammes de taille $n \in \{1..3\}$ ($\{1..3\}$-grammes) lorsque leurs mots en tête et en queue ne sont pas présents dans l'anti-dictionnaire fourni par l'université de Neuchâtel[4] (*IR Multilingual Resources at UniNE*). La sélection des n-grammes est très exhaustive, elle fournit un grand nombre de termes-clés candidats, ce qui permet de maximiser la quantité de candidats présents dans l'ensemble des termes-clés de référence, mais ce qui maximise aussi la quantité de candidats erronés (bruités).

Exemples de $\{1..3\}$-grammes sélectionnés à partir de « bassin moyen du Don » dans la notice d'archéologie de la figure 1 : « bassin », « moyen », « Don », « bassin moyen » et « moyen du Don ».

La **reconnaissance de formes** consiste à sélectionner les unités textuelles qui respectent certains patrons grammaticaux. Les termes-clés candidats sélectionnés par reconnaissance de forme ont l'avantage d'avoir une nature contrôlée avec précision (p. ex. des groupes nominaux), ce qui les rend plus fondés linguistiquement, ainsi que de meilleure qualité que les n-grammes. Dans ce travail, nous utilisons le patron `/(NOM | ADJ)+/` afin de sélectionner les plus longues séquences de noms (noms propres inclus) et d'adjectifs (Hasan & Ng, 2010).

Exemples de `/(NOM | ADJ)+/` sélectionnés à partir de « bassin moyen du Don » dans la notice d'archéologie de la figure 1 : « bassin moyen » et « Don ».

La **sélection de candidats termes** consiste à sélectionner les unités textuelles qui sont potentiellement des termes, tels que définis dans le domaine de l'extraction terminologique. En terminologie, un terme est un mot ou une séquence de mots représentant un concept spécifique à un domaine, ou une discipline. Dans ce travail, nous utilisons l'extracteur terminologique TermSuite (Rocheteau & Daille, 2011), qui est capable de détecter des candidats termes (simples et complexes) ainsi que leurs variantes. Une terminologie est extraite par TermSuite pour chaque corpus (32 119 candidats termes en

4. `http://members.unine.ch/jacques.savoy/clef/index.html`

Archéologie, 16 557 candidats termes en Sciences de l'Information, 21 330 candidats termes en Linguistique, 24 680 candidats termes en Psychologie et 21 020 candidats termes en Chimie) et toutes les entrées de la terminologie apparaissant dans un document de la discipline sont sélectionnés comme termes-clés candidats de ce document [5]. cette terminologie sont extraites comme termes-clés candidats. Contrairement à la méthode de sélection des plus longues séquences de noms et d'adjectifs, la sélection des candidats termes de TermSuite se fonde sur un travail de spécification linguistique et terminologique des termes. Les patrons grammaticaux utilisés par TermSuite sont donc plus précis (p. ex. `/NOM à NOM/`, `/NOM en NOM/`, `/NOM à NOM ADJ/`, etc.) et de longueur plus restreinte puisque les structures à deux ou trois mots lexicaux sont privilégiées.

Exemples de candidats termes sélectionnés à partir de « bassin moyen du Don » dans la notice d'archéologie de la figure 1 : « bassin », « Don », « bassin moyen » et « bassin moyen du Don ».

3.3 Ordonnancement des termes-clés candidats

Un grand nombre de méthodes sont proposées dans la catégorie des méthodes non-supervisées. Parmi elles, les méthodes d'ordonnancement TF-IDF (Spärck Jones, 1972) et TopicRank (Bougouin *et al.*, 2013). De part sa simplicité et sa robustesse, la méthode TF-IDF s'impose comme la méthode de référence pour l'extraction non-supervisée de termes-clés [6], tandis que les méthodes à base de graphe, telles que TopicRank, suscitent un intérêt grandissant. En effet, les graphes permettent de représenter simplement et efficacement les unités textuelles d'un document et leurs relations en son sein. De plus, ils bénéficient de nombreuses études théoriques donnant lieu à des outils et algorithmes efficaces pour résoudre divers problèmes. TF-IDF et TopicRank ont un fonctionnement très différent, ce qui nous permet par la suite d'identifier les facteurs qui influent sur la difficulté de l'extraction automatique de termes-clés.

La méthode **TF-IDF** consiste à extraire en tant que termes-clés les candidats dont les mots sont importants. Un score d'importance (TF-IDF) est attribué à chaque mot des candidats et l'importance d'un candidat est calculé par la somme du score d'importance de ses mots. Selon TF-IDF, un mot est considéré important dans un document s'il y est fréquent (TF élevé) et s'il a une forte spécificité (IDF élevé). Cette dernière est déterminée à partir d'une collection de documents [7] : moins il y a de documents qui contiennent le mot, plus forte est sa spécificité.

TopicRank (Bougouin *et al.*, 2013) extrait les termes-clés qui représentent les sujets les plus importants d'un document. Tout d'abord, TopicRank groupe les termes-clés candidats selon leur appartenance à un sujet, représente les documents sous la forme d'un graphe de sujets et ordonne les sujets selon leur importance dans le graphe (Mihalcea & Tarau, 2004). Enfin, le terme-clé candidat le plus représentatif d'un sujet, celui qui apparaît en premier dans le document, est extrait en tant que terme-clé [8].

TopicRank groupe les termes-clés candidats selon une mesure de similarité lexicale (cf. équation 1). Cependant, TermSuite fournit un groupement terminologique des termes et de leurs variantes. Lorsque les termes-clés candidats sont ceux extraits avec TermSuite, nous tirons profit de ce groupement terme/variantes à la place de celui fondé sur la similarité lexicale. Tenant compte du groupement (moins naïf) de TermSuite, TopicRank distingue alors les candidats « Kostienki 11/II/ » et « Kostienki 21/III/ » qui représentent des faciès différents (cf. figure 1).

$$\text{similarité}(c_1, c_2) = \frac{\|c_1 \cap c_2\|}{\|c_1 \cup c_2\|}, \tag{1}$$

où c_1 et c_2 sont deux termes-clés candidats représentés par des sacs de mots.

5. L'extraction terminologique effectuée depuis tous les documents de chaque collection permet une meilleure précision lors de la détection des variantes des termes. De plus, la taille des collections entre 80 000 et 150 000 mots est faible pour une extraction terminologique, mais ceci est compensé par le haut degré de densité terminologique des collections.
6. Notons qu'une variante de la pondération TF-IDF est utilisée en Recherche d'Information (Robertson *et al.*, 1998; Claveau, 2012, Okapi). Bien que cette variante est jugée plus efficace en Recherche d'Information, celle-ci n'a, à notre connaissance, jamais été employée pour l'extraction automatique de termes-clés. Notre objectif n'étant pas de trouver la meilleure méthode d'extraction de termes-clés, nous utilisons la méthode originale.
7. Dans ce travail, nous utilisons la collection dont est extrait le document.
8. Si nécessaire, les termes-clés extraits sont pondérés et ordonnés selon le score d'importance de leur sujet respectif

4 Expériences

Dans cette section, nous présentons les expériences menées dans le but d'observer l'échelle de difficulté pour l'extraction automatique de termes-clés en domaines de spécialité à partir des méthodes TF-IDF et TopicRank et en fonction des candidats qui sont sélectionnés : $\{1..3\}$-grammes filtrés, plus longues séquences de noms et d'adjectifs et candidats termes.

4.1 Mesure d'évaluation

Afin de mesurer l'échelle de difficulté pour l'extraction automatique de termes-clés en domaines de spécialité, nous utilisons la MAP (*Mean Average Precision*), qui mesure la capacité d'une méthode à ordonner correctement les termes-clés de référence parmi tous les termes-clés candidats, c'est-à-dire à extraire en premier des candidats qui sont présents dans la liste des termes-clés de référence (cf. équation 2). Alors qu'il est plus courant d'utiliser la précision, le rappel et la f-mesures pour comparer les méthodes entre elles, notre choix se porte sur la MAP à cause du nombre variable de termes-clés de référence assignés aux documents par discipline (de 8,0 en linguistique à 16,6 en archéologie). La MAP étant appliquée à tous les candidats ordonnés et non pas à un sous ensemble (p. ex. les 10 premiers, pour la précision, le rappel et la f-mesure), il ne peut y avoir de biais lorsque nous comparons l'extraction de termes-clés entre deux disciplines.

$$\text{MAP} = \frac{1}{\|\text{DOCUMENTS}\|} \sum_{d \in \text{DOCUMENTS}} \frac{\sum_{t_i \in \text{extraction}(d)\ \cap\ \text{référence}(d)} \text{précision@}i}{\|\text{référence}(d)\|} \tag{2}$$

où :
- extraction(d) founie l'ensemble ordonné des termes-clés candidats t_i de rang i pour le document d,
- référence(d) fournie l'ensemble des termes-clés de référence du document d,
- précision@i représente la précision de l'extraction calculée au rang i,
- DOCUMENTS est l'ensemble des documents de la collection pour laquelle les termes-clés sont extraits.

En accord avec l'évaluation menée dans les travaux précédents, nous considérons correcte l'extraction d'une variante flexionnelle d'un terme-clé de référence (Kim *et al.*, 2010). Les opérations de comparaison entre les termes-clés de référence et les termes-clés extraits sont donc effectuées à partir de la racine des mots qui les composent en utilisant la méthode de racinisation de Porter (1980).

4.2 Résultats

La figure 3 montre la performance des méthodes d'extraction automatique de termes-clés lorsque les candidats sélectionnés sont soit les $\{1..3\}$-grammes filtrés, soit les plus longues séquences de noms et d'adjectifs, soit tous les candidats termes extraits par TermSuite (sans filtrage). Notre hypothèse de départ selon laquelle la tâche d'extraction de termes-clés présente un degré de difficulté différent selon la discipline scientifique se vérifie. L'archéologie est la discipline pour laquelle la tâche d'extraction automatique de termes-clés est la moins difficile, la chimie étant la discipline la plus difficile, précédée par la psychologie, les sciences de l'information et la linguistique. Quelle que soit la discipline traitée, nous pouvons aussi observer la faible performance des méthodes d'extraction de termes-clés (cf. exemple figure 4). Ceci peut s'expliquer par le faible rappel maximum pouvant être atteint, ainsi que par l'évaluation stricte qui n'accepte pas les correspondances partielles (p. ex. « articles » et « articles de recherche » qui dans le contexte de la notice de la figure 4 représentent le même concept).

Globalement, les meilleurs résultats sont obtenus avec la méthode TF-IDF. De plus, bien que dans le meilleur cas elle soit compétitive avec TF-IDF, la méthode TopicRank n'est pas stable. Lorsque les $\{1..3\}$-grammes sont utilisés comme candidats nous observons une forte dégradation des résultats de TopicRank, alors que la dégradation des résultats de TF-IDF est plus modérée. Cette différence de comportement face à un ensemble de termes-clés candidats de mauvaise qualité s'explique par le fait que le groupement en sujets de TopicRank n'est pas adapté pour de tels candidats et aussi parce que TF-IDF tire profit de la spécificité des mots (IDF), lui permettant, contrairement à TopicRank, de ne pas attribuer un fort poids aux candidats erronés tels que « d' » (cf. figure 4). En ce qui concerne les résultats obtenus avec les deux autres méthodes de sélection des termes-clés candidats, les performances sont meilleures avec les plus longues séquences de noms et d'adjectifs. La différence de performance observée avec ces deux méthodes de sélection est principalement

liée à la richesse des patrons grammaticaux utilisés par TermSuite. En effet, ses patrons grammaticaux contenant des déterminants et des prépositions ne reflètent qu'une infime quantité de termes-clés de référence (3,5 %) et ont donc pour effet d'ajouter plus de bruit que de candidats positifs.

5 Discussion

À partir des expériences de la section 4, nous constatons la même échelle de difficulté quelque soit la méthode employée (cf. figure 5), ce qui montre que notre hypothèse de départ est valide. Il est toutefois important de noter qu'en observant les statistiques présentées dans le tableau 1, nous pouvons déduire la même échelle de difficulté à partir du rappel maximum. Cependant, le rappel maximum ne peut être obtenu en dehors d'un contexte expérimental. Dans cette section, nous nous fondons sur la nature des collections de données et sur les résultats de l'extraction non-supervisée de termes-clés pour déterminer quels sont les facteurs qui influent sur la difficulté de cette tâche.

Dans un premier temps, nous constatons que la pondération fondée sur la spécificité des mots améliore la stabilité (la robustesse) des méthodes d'extraction de termes-clés qui l'utilisent. Nous en déduisons que la nature linguistique des termes-clés utilisés dans une discipline est un facteur qui influe sur la difficulté de l'extraction des termes-clés. Ainsi, une forte tendance à l'usage de composés syntagmatiques constitués de mots centraux dans la discipline, tels que « réaction » en Chimie (p. ex. « réaction topotactique » et « réaction sonochimique ») ou encore le mot « social », qui est fréquemment utilisé en psychologie (p. ex. « interaction sociale » et « environnement social »), augmente la difficulté de l'extraction des termes-clés.

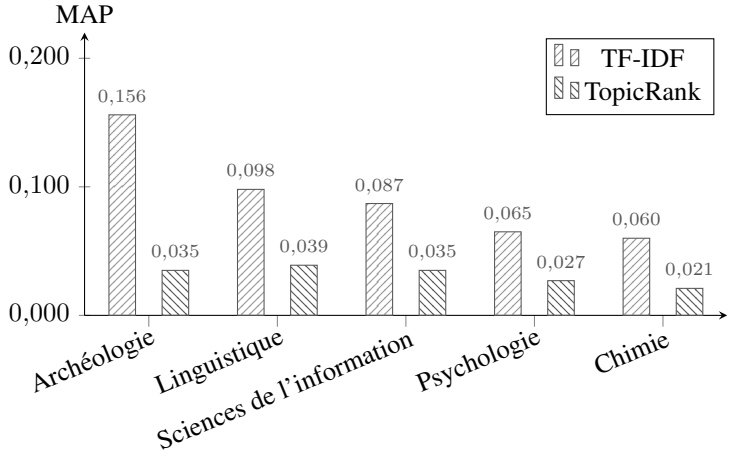
Dans un second temps, nous observons, sauf dans le cas de la psychologie, qu'il y a une correspondance entre l'ordre des disciplines selon la taille des résumés des notices et leur ordre dans l'échelle de difficulté. Ceci s'explique par la façon dont est organisé le discours dans les notices. Si nous prenons les notices d'archéologie, par exemple, celles-ci sont très détaillées. Il est par conséquent aisé d'établir des relations entre les concepts afin de déterminer quels sont ceux les plus importants, à la manière de TopicRank. À l'inverse, les notices de chimie, représentant principalement des comptes rendus d'expériences s'adressent à un lecteur expert pour lequel il est uniquement nécessaire de décrire le contexte expérimental. L'absence de détails dans les notices de certaines disciplines est donc un facteur qui influe sur la difficulté de l'extraction automatique de termes-clés, difficulté qui peut a priori être détectée à partir de la taille des notices.

6 Conclusion et perspectives

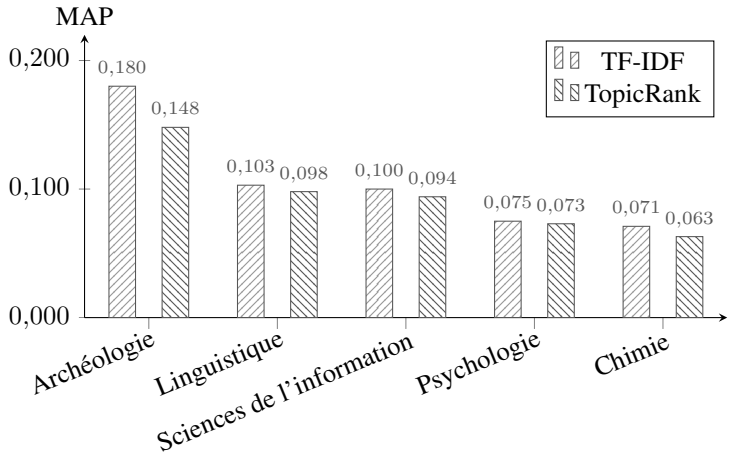
Dans cet article, nous nous intéressons à la tâche d'extraction automatique de termes-clés dans les documents scientifiques et émettons l'hypothèse que sa difficulté est variable selon la discipline des documents traités. Pour vérifier cette hypothèse, nous disposons de notices bibliographiques réparties dans cinq disciplines (archéologie, linguistique, sciences de l'information, psychologie et chimie) auxquelles nous appliquons six systèmes d'extraction automatique de termes-clés différents. En comparant les termes-clés extraits par chaque système avec les termes-clés de référence assignés aux notices dans des conditions réels d'indexation, notre hypothèse se vérifie et nous observons l'échelle suivante (de la discipline la plus facile à la plus difficile) : 1. Archéologie ; 2. Linguistique ; 3. Sciences de l'information ; 4. Psychologie ; 5. Chimie.

À l'issue de nos expériences et de nos observations du contenu des notices, nous constatons deux facteurs ayant un impact sur la difficulté de la tâche d'extraction automatique de termes-clés. Tout d'abord, nous observons que l'organisation du résumé peut aider l'extraction de termes-clés. Un résumé riche en explications et en mises en relations des différents concepts est moins difficile à traiter qu'un résumé énumératif pauvre en explications. Ensuite, le vocabulaire utilisé dans une discipline peut influer sur la difficulté à extraire les termes-clés des documents de cette discipline. Si le vocabulaire spécifique contient des composés syntagmatiques dont certains éléments sont courants dans la discipline, alors il peut être plus difficile d'extraire les termes-clés des documents de cette discipline.

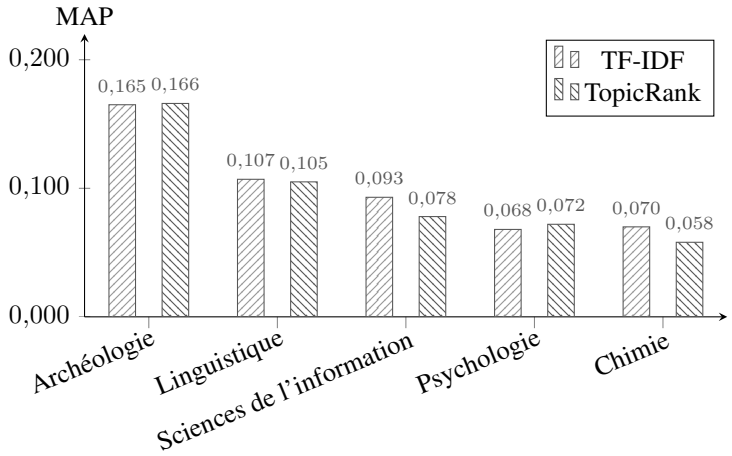
Des deux facteurs identifiés émergent plusieurs perspectives de travaux futurs. Il peut être intéressant d'analyser le discours des documents afin de mesurer, en amont, le degré de difficulté de l'extraction de termes-clés. Avec une telle connaissance, nous pourrions proposer une méthode capable de s'adapter au degré de difficulté en ajustant automatiquement son paramètrage. Cependant, l'analyse que nous proposons dans cet article se fonde uniquement sur le contenu de notices appartenant à cinq disciplines. Il serait pertinent d'étendre cette analyse au contenu intégral des documents scientifiques, ainsi que d'élargir le panel de disciplines utilisées dans ce travail, afin d'établir des catégories de disciplines plus ou moins difficiles à traiter (p. ex. la chimie fait partie des disciplines expérimentales, qui sont difficiles à traiter). Nous



(a) {1..3}-grammes



(b) /(NOM | ADJ)+/



(c) Candidats termes

FIGURE 3 – Performances des méthodes d'extraction de termes-clés en domaines de spécialité à partir de différents type de candidats.

Termes techniques et marqueurs d'argumentation : pour débusquer l'argumentation cachée dans les articles de recherche *Linguistique*

Les articles de recherche présentent les résultats d'une expérience qui modifie l'état de la connaissance dans le domaine concerné. Le lecteur néophyte a tendance à considérer qu'il s'agit d'une simple description et à passer à côté de l'argumentation au cours de laquelle le scientifique cherche à convaincre ses pairs de l'innovation et de l'originalité présentées dans l'article et du bien-fondé de sa démarche tout en respectant la tradition scientifique dans laquelle il s'insère. Ces propriétés spécifiques du discours scientifique peuvent s'avérer un obstacle supplémentaire à la compréhension, surtout lorsqu'il s'agit d'un article en langue étrangère. C'est pourquoi il peut être utile d'incorporer dans l'enseignement des langues de spécialité une sensibilisation aux marqueurs linguistiques (terminologiques et argumentatifs), qui permettent de dépister le développement de cette rhétorique. Les auteurs s'appuient sur deux articles dans le domaine de la microbiologie.

Termes-clés de référence : Langue scientifique*, argumentation*, rhétorique*, langue de spécialité*, enseignement des langues*, linguistique appliquée*, discours scientifique*, article de recherche.

Termes-clés extraits :

{1..3}-grammes

TF-IDF : Argumentation, scientifique, articles, d' argumentation, l' argumentation, tradition scientifique, **discours scientifique**, marqueurs.

TopicRank : Articles, d', qu' il s', débusquer l' argumentation, **articles de recherche**, scientifique, s' agit d', marqueurs d' argumentation.

/(NOM | ADJ)+/

TF-IDF : Argumentation, scientifique, articles, tradition scientifique, **discours scientifique**, marqueurs, microbiologie, domaine.

TopicRank : Article, **argumentation**, recherche, marqueurs, domaine, langue étrangère, scientifique, résultats.

Candidats termes

TF-IDF : Argumentation, scientifique, tradition scientifique, **discours scientifique**, marqueurs, microbiologie, néophyte, marqueurs d' argumentation.

TopicRank : Argumentation, marqueurs, **articles de recherche**, scientifique, techniques, termes, article, langue.

FIGURE 4 – Exemple d'extraction automatique de (huit) termes-clés à partir de la notice de linguistique présentée dans la figure 1. Les termes-clés de référence soulignés sont ceux qui occurrent dans le titre ou le résumé de la notice. Les termes-clés de référence marqués d'une * font partie des termes-clés contrôlés. Les termes-clés extraits mis en gras sont les termes-clés correctement extraits.

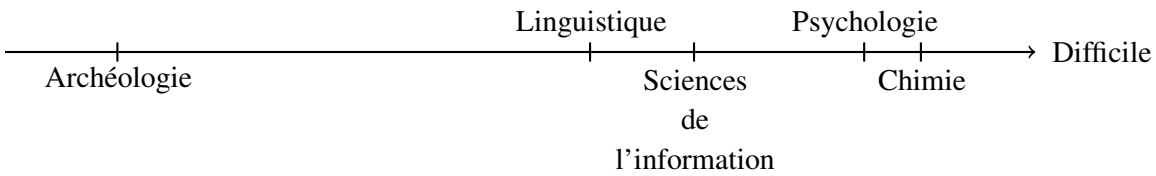


FIGURE 5 – Échelle de difficulté disciplinaire, de la discipline la moins difficile à la discipline la plus difficile à traiter par les méthodes d'extraction automatique de termes-clés.

observons aussi que le vocabulaire utilisé dans une discipline, en particulier celui utilisé pour les termes-clés, peut rendre la tâche d'extraction automatique de termes-clés plus difficile. Il est donc important de bénéficier de resources telles que des thésaurus pour permettre à une méthode d'extraction de termes-clés de s'adapter au domaine. Pour TopicRank, par exemple, avoir connaissance de la terminologie utilisée dans une discipline peut améliorer le choix du terme-clé le plus représentatif d'un sujet. Enfin, il serait intéressant de penser la tâche d'extraction de termes-clés comme une tâche d'extraction d'information pour le remplissage d'un formulaire. En archéologie, par exemple, il pourrait s'agir d'extraire les informations géographiques (pays, régions, etc.), chronologiques (période, culture, etc.), ou encore environnementales (animaux, végétaux, etc.).

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. (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.

CLAVEAU V. (2012). Vectorisation, Okapi et calcul de similarité pour le TAL : pour oublier enfin le TF-IDF (Vectorization, Okapi and Computing Similarity for NLP : Say Goodbye to TF-IDF) [in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, Volume 2 : TALN*, p. 85–98, Grenoble, France : ATALA/AFCP.

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.

D'AVANZO E. & MAGNINI B. (2005). A Keyphrase-Based Approach to Summarization : the LAKE System at DUC-2005. In *Proceedings of DUC 2005 Document Understanding Conference*.

HAN J., KIM T. & CHOI J. (2007). Web Document Clustering by Using Automatic Keyphrase Extraction. In *Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, p. 56–59, Washington, DC, USA : IEEE Computer Society.

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*, p. 365–373, Stroudsburg, PA, USA : Association for Computational Linguistics.

HASAN K. S. & NG V. (2014). Automatic Keyphrase Extraction : A Survey of the State of the Art. In *Proceedings of the Association for Computational Linguistics (ACL)*, Baltimore, Maryland : Association for Computational Linguistics.

HULTH A. (2003). Improved Automatic Keyword Extraction Given More Linguistic Knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, p. 216–223, Stroudsburg, PA, USA : Association for Computational Linguistics.

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, Stroudsburg, PA, USA : Association for Computational Linguistics.

MEDELYAN O. & WITTEN I. H. (2008). Domain-Independent Automatic Keyphrase Indexing with Small Training Sets. *Journal of the American Society for Information Science and Technology*, **59**(7), 1026–1040.

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*, p. 404–411, Barcelona, Spain : Association for Computational Linguistics.

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.

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

ROBERTSON S. E., WALKER STEVE & HANCOCK-BEAULIEU MICHELINE (1998). Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive Track. In *Proceedings of the Text REtrieval Conference (TREC)*, p. 199–210.

ROCHETEAU J. & DAILLE B. (2011). TTC TermSuite - A UIMA Application for Multilingual Terminology Extraction from Comparable Corpora. In *Proceedings of the IJCNLP 2011 System Demonstrations*, p. 9–12, Chiang Mai, Thailand : Asian Federation of Natural Language Processing.

SPÄRCK JONES K. (1972). A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, **28**(1), 11–21.

WITTEN I. H., PAYNTER G. W., FRANK E., GUTWIN C. & NEVILL MANNING C. G. (1999). KEA : Practical Automatic Keyphrase Extraction. In *Proceedings of the 4th ACM Conference on Digital Libraries*, p. 254–255, New York, NY, USA : ACM.