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 la tâche d'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, Sciences de l'Information, Linguistique, Psychologie, 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, which are very useful for researchers. 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 the automatic keyphrase extraction from scientific papers from various areas. Using five corpora representing five areas (Archaeology, Information Sciences, Linguistic, 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 principal 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). Avec l'essor du numérique et, en particulier, le développement des bibliothèques numériques sur Internet, le nombre d'articles scientifiques auxquels les chercheurs ont accès ne cesse d'augmenter. Les termes-clés, le plus souvent assignés par les auteurs, facilitent l'indexation pour la recherche d'articles scientifiques et donnent un bref aperçu de leur contenu. Cependant, de nombreux articles n'ont pas de termes-clés associés et, compte tenu de leur nombre, l'annotation manuelle de ces derniers n'est pas envisageable. Pour pallier ce problème, de plus en plus de chercheurs s'intéressent à l'extraction automatique de termes-clés à partir d'articles scientifiques et certaines campagnes d'évaluation, telles que DEFT (Paroubek *et al.*, 2012) et SemEval (Kim *et al.*, 2010), proposent des tâches sur cette problématique.

L'extraction automatique de termes-clés, ou indexation libre, consiste à extraire les unités textuelles les plus importantes d'un document, en opposition à l'assignation automatique de termes-clés, ou indexation contrôlée, qui consiste à assigner des termes-clés à partir d'un référentiel donné (Paroubek *et al.*, 2012). 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* » à des termes-clés candidats extraits du document. Une collection de documents annotés en termes-clés est nécessaire 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 en fonction de divers indicateurs

comme leur spécificité par rapport au document (Paukkeri & Honkela, 2010) ou les relations de coocurrence que leurs mots entretiennent (Mihalcea & Tarau, 2004). Les méthodes supervisées sont plus performantes que les méthodes non-supervisées, mais leur besoin en données annotées (pour l'apprentissage) pousse les chercheurs à proposer des méthodes non-supervisées compétitives avec les méthodes 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. Les documents de cette nature appartiennent à des disciplines variées, chacune possédant une terminologie qui lui est spécifique. Certains chercheurs s'adaptent à ces documents en réduisant l'ensemble des termes-clés candidats à ceux les plus spécifiques (Kim *et al.*, 2009), ou en tenant compte de la structure des documents pour extraire les termes-clés apparaissant dans les sections les plus favorables (Hofmann *et al.*, 2009). D'autres chercheurs proposent une analyse de ces documents : Shah *et al.* (2003) étudient les sections génériques des articles scientifiques et déterminent celles les plus favorables à l'introduction de termes-clés, tandis que Bertels *et al.* (2012) analysent la sémantique des termes-clés en domaines de spécialité. Dans la continuité des travaux de Shah *et al.* (2003) et de Bertels *et al.* (2012), nous proposons une analyse de l'influence des domaines de spécialité dans l'extraction de termes-clés et faisons l'hypothèse que chaque discipline est traitée avec un degré de difficulté différent. Au moyen de cinq corpus disciplinaires (section 2), nous utilisons différentes méthodes d'extraction automatique de termes-clés (section 3) et observons leur performance en domaine de spécialité pour déduire l'échelle de difficulté disciplinaire (section 4). Enfin, nous proposons une analyse des résultats et déterminons les facteurs qui influent sur cette difficulté (section 5).

2 Collections de données

Nous disposons de cinq collections de notices bibliographiques fournies par l'Institut de l'Information Scientifique et Technique [1] (INIST) : Archéologie, Sciences de l'Information, Linguistique, Psychologie et Chimie. Chacune d'elles contient le titre, le résumé et les termes-clés (de référence) associés à un article. Les termes-clés sont obtenus semi-automatiquement à partir des textes intégraux (non disponibles pour nos travaux) et à partir de ressources disciplinaires telles qu'une terminologie ou des spécifications précises quant aux types d'information que les termes-clés doivent représenter (e.g. lieu, période et autre, en Archéologie).

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 **Sciences de l'Information** contient 706 notices INIST d'articles publiés entre 2001 et 2012 dans six revues différentes (*Documentaliste – Sciences de l'information*, *Document numérique*, 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 **Psychologie** contient 720 notices d'articles publiés 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.).

1. http://www.inist.fr

Statistique	Archéologie	Sciences de l'Information	Linguistique	Psychologie	Chimie
Documents	718	706	716	720	782
Mots/doc.	219,1	119,7	156,4	185,8	104,9
Termes-clés/doc.	17,7	5,8	8,0	11,0	12,9
Mots/terme-clé	1,3	1,7	1,7	1,6	2,2
Termes-clés ne contenant que des Np	41,3 %	13,9 %	7,7 %	9,3 %	6,7 %

TABLE 1 – Caractéristiques des corpus disciplinaires. Le pourcentage de termes-clés ne contenant que des Np correspond au pourcentage de termes-clés de référence qui contiennent uniquement des mots étiquetés Np (nom propre) par l'outil d'étiquetage morphosyntaxique que nous utilisons (Denis & Sagot, 2009, MElt).

Variabilité du gravettien de Kostienki (bassin moyen du Don) et des territoires associés [2] *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 techniques et marqueurs d'argumentation : pour débusquer l'argumentation cachée dans les articles de recherche [3] *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.

Etude d'un condensat acide isocyanurique-urée-formaldéhyde [4] *Chimie*

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

FIGURE 1 – Exemple de notices INIST (dont les termes-clés sont soulignés).

Le tableau 1 présente les caractéristiques des cinq collections de données présentées ci-dessus. 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 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 (e.g. « Paléolithique », « Europe », etc.), tandis qu'en Chimie, nous observons un usage fréquent de notions générales nécessitant une spécialisation presque systématique (e.g. « réaction topotactique », « réaction sonochimique », « réaction électrochimique », etc.). Enfin, il est important de noter la faible proportion de termes-clés apparaissant dans les notices – rappel maximum pouvant être obtenu. Par exemple, dans le corpus de Sciences de l'Information, uniquement 1,3 termes-clés peuvent être extraits des notices parmi les 5,8 associés aux notices, en moyenne.

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 en sont ensuite extraits, 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 d'extraction des candidats et d'ordonnancement de ceux-ci. Intuitivement, l'ordonnancement des candidats

2. http://cat.inist.fr/?aModele=afficheN&cpsidt=17395748
3. http://cat.inist.fr/?aModele=afficheN&cpsidt=20563716
4. http://cat.inist.fr/?aModele=afficheN&cpsidt=6719275

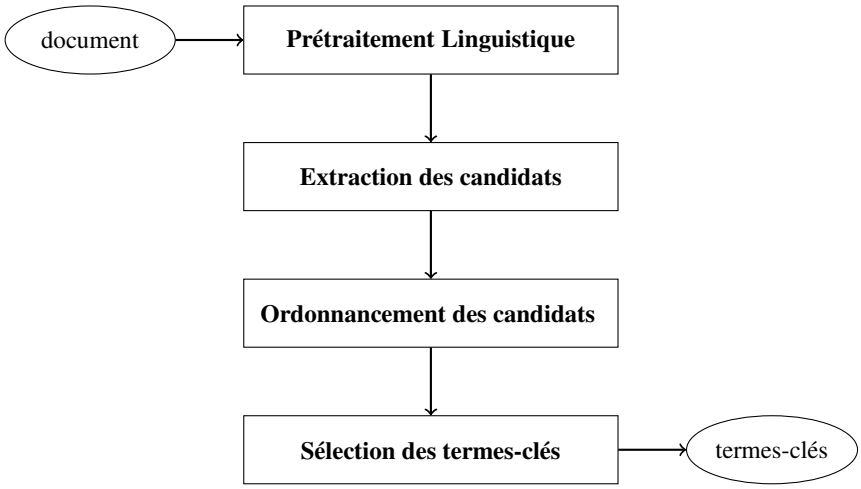


FIGURE 2 – Chaîne de traitements d'un système non-supervisé d'extraction automatique de termes-clés.

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. Nous estimons qu'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, i.e. des candidats qui ne se 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 avec le *PunktSentenceTokenizer* disponible avec la librairie Python NLTK (Bird *et al.*, 2009, *Natural Language ToolKit*) la segmentation en mots est effectuée avec l'outil Bonsai, du Bonsai PCFG-LA parser [5] et l'étiquetage en parties du discours est réalisé avec MElt (Denis & Sagot, 2009). Tous ces outils sont utilisés avec leurs paramètres par défaut.

3.2 Extraction des termes-clés candidats

Dans les travaux précédents, deux approches sont fréquemment utilisées. Soit les candidats sont extraits à partir de n-grammes filtrés, soit ils sont extraits par reconnaissance de formes (Hulth, 2003). Dans ce travail, nous expérimentons trois méthodes différentes : deux méthodes conformes aux approches standards et une méthode utilisant un extracteur terminologique. Un extracteur terminologique fournit des unités textuelles représentant des concepts spécifiques à une discipline. Il semble donc pertinent d'utiliser un tel outil lorsque nous traitons des documents de domaines de spécialité.

L'extraction des **n-grammes** filtrés consiste à extraire toutes les séquences ordonnées de n mots, puis à les filtrer avec une liste de mots outils 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 extrayons les n-grammes de taille $n \in \{1..3\}$ ($\{1..3\}$-grammes) dont les mots en tête et en queue ne sont pas présents dans notre liste de mots outils, fournie par l'université de Neuchâtel [6] (UniNE).

Exemple de $\{1..3\}$-grammes, extraits 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 extraire les unités textuelles qui respectent des patrons définis. Dans ce travail, nous suivons les travaux précédents et extrayons les plus longues séquences de noms communs, de noms propres et d'adjectifs, considérées comme étant les **groupes nominaux** (Hasan & Ng, 2010).

5. http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html
6. http://members.unine.ch/jacques.savoy/clef/index.html

Exemple de groupes nominaux extraits à partir de « [...] (bassin moyen du Don) [...] » dans la notice d'Archéologie de la figure 1 : « bassin moyen » et « Don ».

L'extraction de **candidats termes** consiste à extraire les unités textuelles qui sont potentiellement des termes. 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), capable de détecter des candidats termes (simples et complexes) et leurs variantes. Une terminologie candidate (non filtrée) par corpus est construite automatiquement par TermSuite (32 119 candidats termes en 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 seules les unités textuelles se trouvant dans cette terminologie sont extraites comme termes-clés candidats. Contrairement à la méthode utilisée pour extraire les groupes nominaux, la méthode d'extraction de candidats termes de TermSuite se fonde sur un travail de spécialisation linguistique des termes ayant donné lieu à des patrons plus précis (e.g. `N à N`, `N en N`, etc.).

Exemple de candidats termes extraits 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

Dans la catégorie des méthodes non-supervisées, un grand nombre de méthodes différentes est proposé, dont la méthode TF-IDF (Spärck Jones, 1972) et la méthode TopicRank (Bougouin *et al.*, 2013). De part sa simplicité et sa robustesse, le TF-IDF s'impose comme la méthode de référence [7], tandis que les méthodes à base de graphe, comme TopicRank, suscitent un intérêt grandissant, car les graphes permettent de présenter simplement et efficacement les unités textuelles d'un document et leurs relations en son sein. De plus, les graphes bénéficient de nombreuses études théoriques donnant lieu à des outils et algorithmes efficaces pour résoudre divers problèmes.

La méthode **TF-IDF** consiste à extraire en tant que termes-clés les candidats dont les mots sont importants. Un mot est considéré important dans un document s'il est fréquent dans le document et s'il est spécifique à celui-ci. La spécificité est déterminée à partir d'une collection de documents, de sorte qu'un mot est considéré spécifique lorsqu'il apparaît dans très peu de documents [8].

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, ordonne les sujets selon leur importance dans le graphe, puis sélectionne, pour chacun des meilleurs sujets, son candidat le plus représentatif.

Pour effectuer le groupement en sujets, TopicRank utilise une mesure de similarité lexicale. Cependant, TermSuite fournit un groupement terminologique des termes et des variantes qu'il extrait. 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/ » (voir la figure 1) qui représentent des faciès différents.

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 extrait : $\{1..3\}$-grammes filtrés, groupes nominaux et candidats termes non filtrés.

7. 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 préférons utiliser la méthode originale.

8. Dans ce travail, nous utilisons la collection dont est extrait le document.

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 candidats, i.e. à extraire en premier des candidats qui sont présents dans la liste des termes-clés de référence. Dans notre évaluation, nous considérons correcte l'extraction d'une variante flexionnelle d'un terme-clé de référence. 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 Porter (1980).

4.2 Résultats

La figure 3 montre la performance des méthodes d'extraction de termes-clés lorsque les candidats extraits sont soit des $\{1..3\}$-grammes filtrés, soit des groupes nominaux, soit des candidats termes non filtrés. 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, la Linguistique et les Sciences de l'Information.

Nous observons aussi que le choix des candidats a une forte influence sur certaines méthodes. Avec les $\{1..3\}$-grammes, TopicRank obtient des résultats deux à trois fois inférieurs à ceux obtenus avec les groupes nominaux ou les candidats termes, tandis que les résultats de la méthode TF-IDF subissent des dégradations négligeables. Cela est dû à l'exhaustivité de l'ensemble de $\{1..3\}$-grammes, faisant de celui-ci un ensemble comportant de nombreux candidats non pertinents qui dégradent les performances de TopicRank (dégradation du groupement en sujets, renforcement de liens non pertinents dans le graphe, etc.). Dans le cas de la méthode TF-IDF, cette dégradation est moins conséquente, car les candidats non pertinents ont une faible spécificité et se trouvent donc principalement en queue du classement par importance des candidats (Kim *et al.*, 2009). En opposition, lorsque nous comparons les résultats obtenus à partir des candidats termes à ceux obtenus à partir des groupes nominaux, pour le corpus d'Archéologie, nous observons une légère dégradation des résultats de la méthode TF-IDF et une amélioration de ceux de TopicRank. Pour TF-IDF la dégradation des résultats est due à un ajout important, vis-à-vis des groupes nominaux, de candidats composés de déterminants et de prépositions [9], alors que très peu de termes-clés de référence en contiennent (3,5%). Dans le cas de TopicRank, sa capacité à créer des liens entre des candidats terminologiquement fondés, regroupés terminologiquement et présents dans des documents riches en informations, tels qu'en Archéologie, est un atout important lors de l'extraction de termes-clés.

5 Discussion

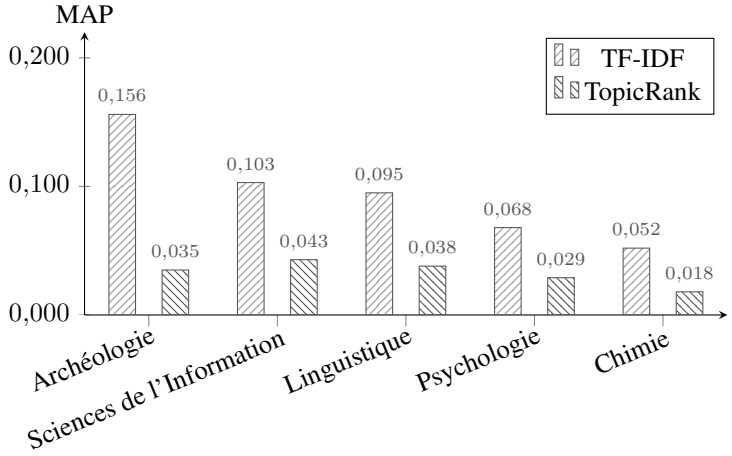
Dans cette section, nous revenons sur les résultats présentés dans la section 4 et pointons, pour les différentes disciplines, les variations qui, selon nous, influent sur la difficulté de la tâche d'extraction de termes-clés en domaines de spécialité. À partir des résultats obtenus, nous déduisons l'échelle de difficulté suivante (de la discipline la plus difficile à la plus facile) :

1. Chimie
2. Psychologie
3. Linguistique
4. Sciences de l'Information
5. Archéologie

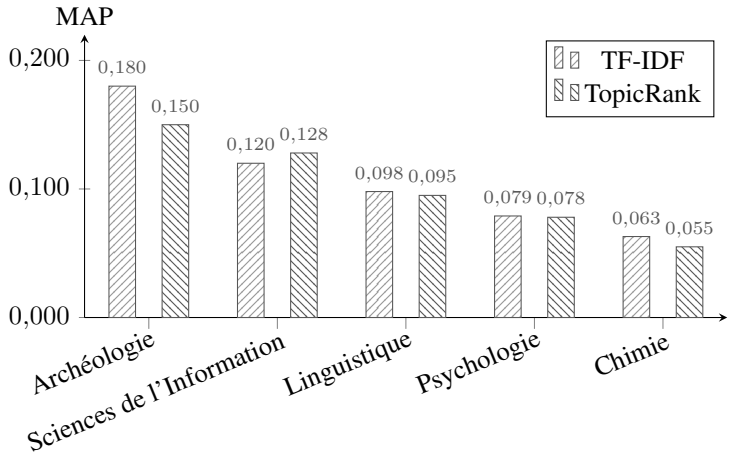
Selon cette échelle de difficulté, ainsi que selon nos observations du contenu des notices, nous définissons trois catégories pour lesquelles la difficulté n'est pas la même :

1. Travaux expérimentaux (Chimie)
2. Travaux analytiques (Psychologie, Linguistique et Sciences de l'Information)
3. Travaux pratiques, i.e. fondés sur des faits non sujets à subjectivité (Archéologie)

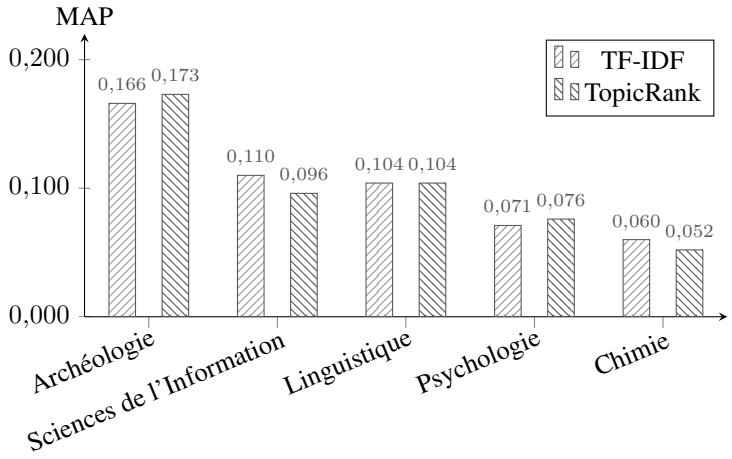
9. Les candidats termes extraits sont environ deux fois plus nombreux que les groupes nominaux et la moitié d'entre eux contiennent des prépositions et des déterminants.



(a) {1..3}-grammes



(b) Groupes nominaux



(c) Candidats termes

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

Dans un premier temps, nous constatons que l'usage d'une mesure de spécificité peut améliorer la performance des méthodes d'extraction de termes-clés. Nous en déduisons que la nature linguistique des termes utilisés dans une discipline est un facteur influant sur la difficulté de l'extraction des termes-clés pour cette même discipline. Ainsi, une forte tendance à l'usage de composés syntagmatiques constitués de mots généraux dans la discipline, tels que « réaction » qui est présent dans le terme-clé « réaction topotactique » en Chimie, augmente la difficulté de l'extraction des termes-clés.

Dans un second temps, nous constatons que la capacité à créer des liens entre différentes unités textuelles peut aider lors de l'extraction des termes-clés. Après observation du contenu des notices, nous remarquons que l'organisation du discours du résumé dans les différentes disciplines est un second facteur influant sur la difficulté de la tâche d'extraction de termes-clés. Pour chaque discipline, le lecteur visé n'est pas le même et le résumé est donc organisé différemment. Dans le cas de documents se basant sur des faits concrets, tels que les documents d'Archéologie, le lecteur (archéologue ou non) a besoin d'une définition du contexte et des relations entre les faits donnés. Un document insistant sur différents éléments importants et créant des liens entre ces éléments est plus aisé à traiter qu'un document se reposant sur un acquis supposé (non explicité) du lecteur. Une observation similaire peut-être faite pour les travaux analytiques, où les hypothèses sont clairement explicitées. En contradiction, les documents au sujet de travaux expérimentaux sont très techniques et se reposent sur un acquis supposé du lecteur. En Chimie, les notices sont très souvent énumératives et dépourvues de détails explicatifs, superflus pour un initié. Dans ce cas, moins de liens sont établis entre les termes-clés candidats, et la tâche d'extraction automatique de termes-clés est plus difficile.

6 Conclusion et perspectives

Dans cet article nous étudions la difficulté de la tâche d'extraction automatique de termes-clés appliquée à différents domaines scientifiques. Nous utilisons cinq collections disciplinaires de notices bibliographiques annotées en termes-clés dans des conditions réelles d'indexation. Pour l'Archéologie, les Sciences de l'Information, la Linguistique, la Psychologie et la Chimie nous observons l'échelle de difficulté suivante (de la discipline la plus difficile à la plus facile) : 1. Chimie ; 2. Psychologie ; 3. Linguistique ; 4. Sciences de l'Information ; 5. Archéologie ;

À l'issue de nos expériences et de nos observations à partir 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, le vocabulaire utilisé dans une discipline peut influer sur la difficulté à extraire des termes-clés à partir de documents de cette discipline. Si le vocabulaire spécifique contient des composés syntagmatiques dont certains éléments sont généraux dans la discipline, alors il peut être plus difficile d'extraire des termes-clés de documents de cette discipline. Ensuite, 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 éléments est moins difficile à traiter qu'un résumé énumératif.

Des deux facteurs identifiés émergent deux perspectives de travaux futurs. La nature différente des unités textuelles définies comme termes-clés selon les disciplines implique un besoin d'adapter l'extraction des candidats à la discipline traitée. L'une des ressources disciplinaires utilisées par les indexeurs professionnels étant la spécification précise du type d'information que doivent représenter les termes-clés, il serait intéressant d'utiliser une méthode d'extraction d'information à partir de formulaires (e.g. remplissage de champs périodes et lieus, en Archéologie). Une autre perspective serait d'analyser le discours afin de mesurer, en amont, le degré de difficulté de l'extraction des termes-clés. Avec une telle connaissance, nous pourrions proposer une méthode capable de s'adapter au degré de difficulté d'un document, en ajustant automatiquement ses différents paramètres. Néanmoins, l'analyse que nous proposons dans cet article se fonde uniquement sur le contenu de notices. Il serait pertinent d'étendre cette analyse faite à partir de résumés à une analyse faite à partir de textes intégraux.

Remerciements

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

Références

BERTELS A., DE HERTOG D. & HEYLEN K. (2012). Étude sémantique des mots-clés et des marqueurs lexicaux

stables dans un corpus technique (Semantic Analysis of Keywords and Stable Lexical Markers in a Technical Corpus) [in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN*, p. 239–252, Grenoble, France : ATALA/AFCP.

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.

HOFMANN K., TSAGKIAS M., MEIJ E. & DE RIJKE M. (2009). The Impact of Document Structure on Keyphrase Extraction. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, p. 1725–1728 : ACM.

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., KAN M.-Y. & BALDWIN T. (2009). An Unsupervised Approach to Domain-Specific Term Extraction. In *Proceedings of the 2009 Australasian Language Technology Association Workshop*.

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.

PAUKKERI M.-S. & HONKELA T. (2010). Likey : Unsupervised Language-Independent Keyphrase Extraction. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, p. 162–165, Stroudsburg, PA, USA : Association for Computational Linguistics.

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.

SHAH P. K., PEREZ IRATXETA C., BORK P. & ANDRADE M. A. (2003). Information Extraction from Full Text Scientific Articles : Where are the Keywords ? *BMC bioinformatics*, **4**(1), 20.

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.