HAL open science

État de l'art des méthodes d'extraction automatique de termes-clés

Adrien Bougouin

To cite this version:

Adrien Bougouin. État de l'art des méthodes d'extraction automatique de termes-clés. Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL), Jun 2013, Sables d'Olonne, France. hal-00821671

HAL Id: hal-00821671

https://hal.science/hal-00821671v1

Submitted on 11 May 2013

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.

État de l'art des méthodes d'extraction automatique de termes-clés

Adrien Bougouin
LINA - UMR CNRS 6241, Université de Nantes, France
adrien.bougouin@univ-nantes.fr

RÉSUMÉ

Cet article présente les principales méthodes d'extraction automatique de termes-clés. La tâche d'extraction automatique de termes-clés consiste à analyser un document pour en extraire les expressions (phrasèmes) les plus représentatives de celui-ci. Les méthodes d'extraction automatique de termes-clés sont réparties en deux catégories : les méthodes supervisées et les méthodes non supervisées. Les méthodes supervisées réduisent la tâche d'extraction de termes-clés à une tâche de classification binaire (tous les phrasèmes sont classés parmi les termes-clés ou les non termes-clés). Cette classification est possible grâce à une phase préliminaire d'apprentissage, phase qui n'est pas requise par les méthodes non-supervisées. Ces dernières utilisent des caractéristiques (traits) extraites du document analysé (et parfois d'une collection de documents de références) pour vérifier des propriétés permettant d'identifier ses termes-clés.

ABSTRACT

State of the Art of Automatic Keyphrase Extraction Methods

This article presents the state of the art of the automatic keyphrase extraction methods. The aim of the automatic keyphrase extraction task is to extract the most representative terms of a document. Automatic keyphrase extraction methods can be divided into two categories : supervised methods and unsupervised methods. For supervised methods, the task is reduced to a binary classification where terms are classified as keyphrases or non keyphrases. This classification requires a learning step which is not required by unsupervised methods. The unsupervised methods use features extracted from the analysed document (sometimes a document collection) to check properties which allow keyphrase identification.

MOTS-CLÉS : extraction de termes-clés ; méthodes supervisées ; méthodes non-supervisées ; état de l'art .

KEYWORDS: keyphrase extraction ; supervised methods ; unsupervised methods ; state of the art .

1 Introduction

Les termes-clés sont des mots ou des expressions (multi-mots) représentant les aspects principaux qui sont abordés dans un document. De ce fait, ils sont utilisés dans de nombreux domaines du Traitement Automatique des Langues (TAL). Turney (1999) émet l'hypothèse qu'ils peuvent faciliter la lecture d'un utilisateur en lui permettant de surfer d'un point clé à un autre lorsqu'ils

sont mis en évidence dans un texte. D'autres chercheurs utilisent leurs vertus synthétiques dans des méthodes de construction automatique de résumés (Wan *et al.*, 2007; Litvak et Last, 2008; Boudin et Morin, 2013), mais ils s'avèrent surtout de plus en plus utiles avec l'essor de l'Internet et la disponibilité de nombreux documents numériques qu'il faut pouvoir indexer de manière pertinente pour faciliter leur recherche par des utilisateurs (Medelyan et Witten, 2008). Dans ce contexte de recherche d'information, les termes-clés peuvent aussi être directement bénéfiques aux utilisateurs en servant de suggestions à une requête qu'ils essaient de formuler (Jones et Staveley, 1999).

Bien que les termes-clés soient utiles pour de multiples tâches, très peu de documents en sont pourvus, du fait du coût important de production de ceux-ci, en termes de temps et de ressources humaines. Pour y remédier de nombreux chercheurs s'intéressent à l'extraction automatique de ceux-ci et certaines campagnes d'évaluations, telles que DEFT (Paroubek *et al.*, 2012) et SemEval (Kim *et al.*, 2010), proposent des tâches d'extraction automatique de termes-clés dans le but de confronter les différents systèmes existants. Pour ce faire, les données et la méthode d'évaluation sont les mêmes pour tous les systèmes.

Il existe aussi une autre tâche nommée assignation automatique de termes-clés. Cette tâche est très proche de l'extraction automatique de termes-clés, mais elle est plus contrôlée. Elle consiste aussi à donner un ensemble de termes-clés pour un document, mais certains de ces termes peuvent ne pas être présents dans celui-ci. Ceci est dû au fait que les méthodes d'assignation de termes-clés utilisent des ressources supplémentaires telles que des référentiels terminologiques. Ceux-ci contiennent des termes spécifiques au(x) domaine(s) traité(s) et l'assignation de ces termes peut être déclenchée par la présence de certains autres dans le document analysé.

Dans cet article, seules les méthodes d'extraction automatique de termes-clés sont présentées. Celles-ci appartiennent à deux catégories distinctes : les méthodes supervisées et les méthodes non-supervisées. Dans le cas supervisé, l'extraction des termes-clés est effectuée grâce à un apprentissage préalable servant à calibrer la méthode avec un corpus dont les documents sont annotés en termes-clés. Les méthodes non-supervisées ne requièrent pas de phase d'apprentissage. Elles exploitent des représentations efficaces des documents ainsi que des propriétés définies à partir de traits statistiques pour extraire les termes-clés parmi des termes candidats.

Dans la section 2 de cet article, nous présentons les méthodes existantes d'extraction automatique de termes-clés, en commençant par les méthodes non-supervisées, puis les méthodes supervisées. Dans la section 3 nous terminons par un bilan de l'état de l'art et nous discutons des perspectives de travaux futurs.

2 Les méthodes d'extraction automatique de termes-clés

L'extraction de termes-clés est une tâche qui consiste à analyser un document et à en extraire les aspects importants. Alors que les méthodes de résumé automatique utilisent des phrases pour construire une vision synthétique du document, l'extraction de termes-clés se focalise sur les unités textuelles qui composent ces phrases. Un ensemble de termes-clés peut donc être perçu comme un résumé dont les points clés sont exprimés sans liaisons entre eux. Les unités textuelles sur lesquelles travaillent les systèmes d'extraction automatique de termes-clés sont appelées termes candidats. Ces derniers sont des mots ou des multi-mots (phrasèmes) pouvant

être promus au statut de terme-clé.

L'extraction de termes candidats est une étape préliminaire de l'extraction de termes-clés, que ce soit pour les méthodes non-supervisées ou supervisées. Cette étape est importante, car si certains termes-clés du document analysé ne sont pas présents dans l'ensemble des termes candidats, alors ceux-ci ne pourront pas être extrait. Hulth (2003) étudie trois méthodes d'extraction des termes candidats. L'une consiste à extraire les chunks nominaux [1], tandis que les deux autres extraient tous les n-grammes et les filtrent, soit pour retirer les termes contenant des mots outils dans le premier cas, soit pour ne retenir que les termes respectant certains patrons syntaxiques dans le second cas (usage des parties du discours). Dans ses expériences Hulth (2003) montre que l'extraction de termes-clés à partir de n-grammes filtrés avec les mots outils donne les meilleurs résultats parmi les trois méthodes qu'elle propose.

Les travaux de Hulth (2003) sont évalués avec un corpus dont les documents sont des résumés d'articles scientifiques. Cependant, dans d'autres domaines tels que la bio-médecine, la nature des termes à extraire n'est pas la même. En effet, ce sont les acronymes et les entités nommées (noms de protéines par exemple) qu'il est nécessaire d'extraire en tant que termes-clés (Nobata *et al.*, 2008). Pour cela, l'extraction de termes candidats est spécifique au domaine d'application. Les méthodes d'extraction de termes-clés présentées dans cet article traitent des documents supposés sans spécificités particulières, les méthodes d'extraction de termes candidats sont donc les mêmes que celles expérimentées par Hulth (2003), mais il est envisageable de les adapter à des domaines présentant des spécificités particulières.

Utilisés avec les méthodes non-supervisées, les termes candidats sont ordonnés selon un score d'importance obtenu soit à partir d'eux-mêmes, soit à partir de l'importance des mots qui les composent. Si une méthode s'appuie uniquement sur les mots, alors le score d'un terme candidat est généralement calculé en faisant la somme des mots qui le composent. Cependant, ceci n'est pas toujours juste, c'est donc un inconvénient important des méthodes travaillant sur les mots pour extraire les termes-clés. En effet, la sommation peut privilégier des termes qui contiennent beaucoup de mots non-importants vis-à-vis de termes contenant très peu de mots, mais importants.

Utilisés dans les méthodes supervisées, les termes candidats sont classés en tant que termes-clés ou non termes-clés grâce à des méthodes de classification.

2.1 Méthodes non-supervisées

Les méthodes non-supervisées d'extraction de termes-clés ont la particularité de s'abstraire du domaine et de la langue des documents à analyser [2]. Cette abstraction est due au fait que les termes candidats sont analysés avec des règles simples déduites à partir de traits statistiques issus seulement du texte analysé, ou bien d'un corpus de référence non annoté.

De nombreuses approches sont proposées. Certaines se fondent uniquement sur des statistiques alors que d'autres les combinent avec des représentations plus complexes des documents. Ces

1. Un chunk est une unité minimale de sens constituée d'un ou de plusieurs mots. Un chunk nominal est un chunk dont la tête est un nom ou un pronom. Par exemple, dans « Nous avons une bonne politique qualitative. », « Nous » et « une bonne politique qualitative » sont des chunks nominaux.

2. L'abstraction de la langue est vraie pour ce qui est de la méthodologie, cependant les pré-traitements tels que la segmentation en phrases, en mots et l'étiquetage en parties du discours sont eux spécifiques à la langue.

représentations peuvent aller de groupes de mots sémantiquement similaires à des graphes dont les nœuds sont des unités textuelles (mots, expressions, phrases, etc.) liées par des relations de recommandation [3].

2.1.1 Approches statistiques

Plusieurs approches cherchent à définir ce qu'est un terme-clé en s'appuyant sur certains traits statistiques et en étudiant leur rapport avec la notion d'importance d'un terme candidat. Plus un terme candidat est jugé important vis-à-vis du document analysé, plus celui-ci est pertinent en tant que terme-clé.

TF-IDF (cf. équation 1) de Jones (1972) et Likey (cf. équation 2) de Paukkeri et Honkela (2010) sont deux méthodes qui comparent le comportement d'un terme candidat dans le document analysé avec son comportement dans une collection de documents (corpus de référence). L'objectif est de trouver les termes candidats dont le comportement dans le document varie positivement comparé à leur comportement global dans la collection. Dans les deux méthodes ceci s'exprime par le fait qu'un terme a une forte importance vis-à-vis du document analysé s'il y est très présent, alors qu'il ne l'est pas dans le reste de la collection.

$$\text{TF-IDF}(\text{terme}) = TF(\text{terme}) \times \log\left(\frac{N}{DF(\text{terme})}\right) \tag{1}$$

$$Likey(\text{terme}) = \frac{\text{rang}_{\text{document}}(\text{terme})}{\text{rang}_{\text{corpus}}(\text{terme})} \tag{2}$$

Dans TF-IDF, TF représente le nombre d'occurrences d'un terme dans le document analysé et DF représente le nombre de documents dans lequel il est présent, N étant le nombre total de documents. Plus le score TF-IDF d'un terme candidat est élevé, plus celui-ci est important dans le document analysé. Dans Likey, le rang d'un terme candidat dans le document et dans le corpus est obtenu à partir de son nombre d'occurrences, respectivement dans le document et dans le corpus de référence. Plus le rapport entre ces deux rangs est faible, plus le terme candidat évalué est important dans le document analysé.

Okapi (ou BM25) (Robertson *et al.*, 1999) est une mesure alternative à TF-IDF. En Recherche d'Information (RI), celle-ci est plus utilisée que le TF-IDF. Bien que l'extraction automatique de termes-clés soit une discipline à la frontière entre le TAL et la RI, la méthode de pondération Okapi n'a, à notre connaissance, pas été appliquée pour l'extraction de termes-clés. Dans l'article de Claveau (2012), Okapi est décrit comme un TF-IDF prenant mieux en compte la longueur des documents. Cette dernière est utilisée pour normaliser le TF (qui devient TF_{BM25}) :

$$\text{Okapi}(\text{terme}) = TF_{BM25}(\text{terme}) \times \log\left(\frac{N - DF(\text{terme}) + 0,5}{DF(\text{terme}) + 0,5}\right) \tag{3}$$

$$TF_{BM25} = \frac{TF(\text{terme}) \times (k_1 + 1)}{TF(\text{terme}) + k_1 \times \left(1 - b + b \times \frac{DL}{DL_{\text{moyenne}}}\right)} \tag{4}$$

3. Pour une étude comparative de certaines des méthodes par regroupement (Liu *et al.*, 2009) et à base de graphe (Mihalcea et Tarau, 2004; Wan et Xiao, 2008b), voir l'article de Hasan et Ng (2010).

Dans la formule (4), k_1 et b sont des constantes fixées à 2 et $0,75$ respectivement. DL représente la longueur du document analysé et $DL_{moyenne}$ la longueur moyenne des documents de la collection utilisée.

Barker et Cornacchia (2000) estiment que les grands phrasèmes sont plus informatifs et qu'ils doivent être privilégiés. Pour cela, leur approche est très simple : plus un groupe nominal est long et fréquent dans le document analysé, plus il est jugé pertinent en tant que terme-clé de ce document. Cependant, pour éviter la répétition dans le texte, les auteurs des documents utilisent les même expression sous des formes alternatives (plus courtes, par exemple). La fréquence d'une expression ne reflète donc pas forcément sa fréquence réelle d'utilisation, car celle-ci est répartie dans les différentes alternatives. De ce fait, Barker et Cornacchia (2000) repèrent dans les groupes nominaux la tête nominale et utilisent en plus la fréquence de celle-ci.

Tomokiyo et Hurst (2003) tentent de vérifier deux propriétés, en utilisant des modèles de langue uni-grammes et n-grammes et en calculant leur divergence (Kullback-Leibler). Les deux propriétés qu'ils tentent de vérifier sont les suivantes :

- La grammaticalité : un terme-clé doit être bien formé syntaxiquement.
- L'informativité : un terme-clé doit capturer au moins une des idées essentielles exprimées dans le document analysé.

Pour un terme candidat donné, plus sa probabilité en passant du modèle uni-gramme généré à partir du document vers le modèle n-gramme généré à partir du même document augmente, plus il respecte la propriété de grammaticalité. De même, plus sa probabilité en passant du modèle n-gramme généré à partir d'un corpus de référence vers le modèle n-gramme généré à partir du document analysé augmente, plus le terme candidat est informatif.

La méthode que propose Ding *et al.* (2011) utilise TF-IDF comme indicateur de l'importance d'un terme-clé. Dans un ensemble, cette importance doit être maximisée pour chaque terme-clé, mais les auteurs estiment que ceci n'est pas suffisant. Comme Tomokiyo et Hurst (2003), ils définissent deux propriétés qui doivent être respectées :

- La couverture : un ensemble de termes-clés doit couvrir l'intégralité des sujets abordés dans le document représenté.
- La cohérence : les termes-clés doivent être cohérents entre eux.

La propriété de couverture est évaluée avec le modèle *Latent Dirichlet Allocation* (LDA) qui donne la probabilité d'un terme candidat sachant un sujet. La cohérence est évaluée pour chaque paire de termes-clés de l'ensemble avec la mesure d'information mutuelle. Ces deux propriétés sont définies comme contraintes que les auteurs utilisent avec une méthode de programmation par les entiers (technique d'optimisation), la maximisation de la pertinence de chaque terme-clé étant l'objectif à atteindre.

Les traits statistiques utilisés dans les méthodes précédentes sont uniquement utilisés pour déterminer un score de pertinence des termes candidats en tant que termes-clés. Une donnée statistique non citée précédemment, mais pourtant récurrente dans les méthodes d'extraction de termes-clés, est la fréquence de co-occurrences entre deux phrasèmes (termes). Deux phrasèmes co-occurrent s'ils apparaissent ensemble dans le même contexte. La co-occurrence peut être calculée de manière stricte (les phrasèmes doivent être côte-à-côte) ou bien dans une fenêtre de mots. Compter le nombre de co-occurrences entre deux termes permet d'estimer s'ils sont sémantiquement liés ou non. Ce lien sémantique à lui seul ne peut pas servir à extraire des

termes-clés, mais il permet de mieux organiser les termes d'un document pour affiner l'extraction (Matsuo et Ishizuka, 2004; Liu *et al.*, 2009; Mihalcea et Tarau, 2004).

2.1.2 Approches par regroupement

L'objectif des approches par regroupement est de définir des groupes dont les unités textuelles partagent une ou plusieurs caractéristiques communes. Ainsi, lorsque des termes-clés sont extraits à partir de chaque groupe, cela permet de mieux couvrir le document analysé selon les caractéristiques utilisées.

Dans la méthode de Matsuo et Ishizuka (2004), ce sont les termes (phrasèmes) qui sont regroupés. Parmi ceux-ci, seuls les plus fréquents sont concernés par le regroupement. Celui-ci s'effectue en fonction du lien sémantique [4] entre les termes. Après le regroupement, la méthode consiste à comparer les termes candidats du document analysé avec les groupes de termes fréquents, en faisant l'hypothèse qu'un terme candidat qui co-occurre plus que selon toute probabilité avec les termes fréquents d'un ou plusieurs groupes est plus vraisemblablement un terme-clé.

Dans l'algorithme KeyCluster, Liu *et al.* (2009) utilisent aussi un regroupement sémantique, mais dans leur cas ils considèrent les mots du document analysé et ils excluent les mots outils. Dans chaque groupe sémantique, le mot qui est le plus proche du centroïde est sélectionné comme mot de référence. L'ensemble des mots de référence est ensuite utilisé pour filtrer les termes candidats en ne considérant comme termes-clés que ceux qui contiennent au moins un mot de référence (tous les mots de référence devant être utilisés dans au moins un terme-clé).

2.1.3 Approches à base de graphe

Les approches à base de graphe consistent à représenter le contenu d'un document sous la forme d'un graphe. La méthodologie appliquée est issue de PageRank (Brin et Page, 1998), un algorithme d'ordonnancement de pages Web (nœuds du graphe) grâce aux liens de recommandation qui existent entre elles (arcs du graphe). TextRank (Mihalcea et Tarau, 2004) et SingleRank (Wan et Xiao, 2008b) sont les deux adaptations de base de PageRank pour l'extraction automatique de termes-clés [5]. Dans celles-ci, les pages Web sont remplacées par des unités textuelles dont la granularité est le mot et un arc est créé entre deux nœuds si les mots qu'ils représentent co-occurrent dans une fenêtre de mots donnée.

Le graphe est noté $G = (N, A)$, où N est l'ensemble des nœuds du graphe et où A est l'ensemble de ses arcs entrants et sortant : $A_{entrant} \cup A_{sortant}$ [6]. Pour chaque nœud du graphe, un score est calculé par un processus itératif destiné à simuler la notion de recommandation d'une unité textuelle par d'autres [7] (cf. équation 5). Ce score à chaque nœud n_i permet d'ordonner les mots par degré d'importance dans le document analysé. La liste ordonnée des mots peut ensuite être

4. Deux phrasèmes qui co-occurrent fréquemment ensemble sont jugés sémantiquement liés.
5. TextRank a aussi été utilisé pour faire du résumé automatique.
6. Dans le cas de TextRank et de SingleRank $A_{entrant} = A_{sortant}$, car le graphe n'est pas orienté.
7. Plus le score d'une unité textuelle est élevé, plus celle-ci est importante dans le document analysé.

utilisée pour extraire les termes-clés.

$$S(n_i) = (1-\lambda) + \lambda \times \sum_{n_j \in A_{\text{entrant}}(n_i)} \frac{p_{j,i} \times S(n_j)}{\sum_{n_k \in A_{\text{sortant}}(n_j)} p_{j,k}} \tag{5}$$

λ est un facteur d'atténuation qui peut être considéré ici comme la probabilité pour que le nœud n_i soit atteint par recommandation. $p_{j,i}$ représente le poids de l'arc allant du nœud n_j vers le nœud n_i, soit le nombre de co-occurrences entre les deux mots i et j [8].

Dans leurs travaux, Wan et Xiao (2008b) s'intéressent à l'ajout d'informations dans le graphe grâce à des documents similaires (voisins) et aux relations de co-occurrences qu'ils possèdent (ExpandRank). L'objectif est de faire mieux ressortir les mots importants du graphe en ajoutant de nouveaux liens de recommandation ou bien en renforçant ceux qui existent déjà. L'usage de documents similaires peut cependant ajouter ou renforcer des liens qui ne devraient pas l'être. Pour éviter cela, les auteurs réduisent l'impact des documents voisins en utilisant leur degré de similarité avec le document analysé. Une alternative à ExpandRank, CollabRank, également proposée par Wan et Xiao (2008a), fonctionne de la même manière, mais certains choix des auteurs rendent impossible l'usage du degré de similarité pour réduire l'impact des documents voisins. Les résultats moins concluants de CollabRank tendent à confirmer l'importance de l'usage du degré de similarité.

Dans l'optique d'améliorer encore TextRank/SingleRank, Liu *et al.* (2010) proposent une méthode qui cherche cette fois-ci à augmenter la couverture de l'ensemble des termes-clés extraits dans le document analysé (TopicalPageRank). Pour ce faire, ils tentent d'affiner le rang d'importance des mots dans le document en tenant compte de leur rang dans chaque sujet abordé. Le rang d'un mot pour un sujet est obtenu en intégrant à son score PageRank la probabilité qu'il appartienne au sujet (cf. équation 6). Le rang global d'un terme candidat est ensuite obtenu en fusionnant ses rangs pour chaque sujet.

$$S_{\text{sujet}}(N_i) = (1-\lambda) \times p(\text{sujet}|i) + \lambda \times \sum_{N_j \in A_{\text{entrant}}(N_i)} \frac{p_{j,i} \times S(N_j)}{\sum_{N_k \in A_{\text{sortant}}(N_j)} p_{j,k}} \tag{6}$$

Les approches à bases de graphe présentées ci-dessus effectuent toutes un ordonnancement des mots du document analysé selon leur importance dans celui-ci. Pour extraire les termes-clés il est donc nécessaire d'effectuer du travail supplémentaire à partir de la liste ordonnée de mots. Dans la méthode TextRank, les k mots les plus importants sont sélectionnés et retournés (après que ceux apparaissant en collocation dans le document aient été concaténés). La technique utilisée dans les autres méthodes consiste à ordonner les termes candidats en fonction de la somme du score des mots qui les composent. Cependant, puisque l'un des avantages du graphe est que les nœuds peuvent avoir une granularité contrôlée, Liang *et al.* (2009) décident d'utiliser des mots et des multi-mots au lieu de simples mots et de tirer profit de traits supplémentaires, la taille du terme ou encore sa première position dans le document analysé.

8. TextRank utilise un graphe non-pondéré. Dans ce cas, $p_{j,i}$ vaut toujours 1.

2.2 Méthodes supervisées

Les méthodes supervisées sont des méthodes capables d'apprendre à réaliser une tâche particulière, soit ici l'extraction de termes-clés. L'apprentissage se fait grâce à un corpus dont les documents sont annotés en termes-clés. L'annotation permet d'extraire les exemples et les contres-exemples dont les traits statistiques et/ou linguistiques servent à apprendre une classification binaire. La classification binaire consiste à indiquer si un terme candidat est un terme-clé ou non.

De nombreux algorithmes d'apprentissage sont utilisés dans divers domaines. Ils peuvent potentiellement s'adapter à n'importe quelle tâche, dont celle de l'extraction automatique de termes-clés. Les algorithmes utilisés pour celle-ci construisent des modèles probabilistes, des arbres de décision, des Séparateurs à Large Marge (SVM) ou encore des réseaux de neurones [9].

KEA (Witten *et al.*, 1999) est une méthode qui utilise une classification naïve bayésienne pour attribuer un score de vraisemblance à chaque terme candidat, le but étant d'indiquer s'ils sont des termes-clés ou non [10]. Witten *et al.* (1999) utilisent trois distributions conditionnelles apprises à partir du corpus d'apprentissage. La première correspond à la probabilité pour que chaque terme candidat soit étiqueté *oui* (terme-clé) ou *non* (non terme-clé). Les deux autres correspondent à deux différents traits qui sont le poids TF-IDF du terme candidat et sa première position dans le document :

$$P(\text{terme}) = \frac{P_{\text{oui}}(\text{terme})}{P_{\text{oui}}(\text{terme}) + P_{\text{non}}(\text{terme})} \tag{7}$$

$$P_{\text{oui}}(\text{terme}) = P(\text{terme}|\text{oui}) \times \prod_{\text{trait}\in\{\text{TF-IDF,position}\}} P_{\text{trait}}\left(\text{trait}(\text{terme})|\text{oui}\right)$$

$$P_{\text{non}}(\text{terme}) = P(\text{terme}|\text{non}) \times \prod_{\text{trait}\in\{\text{TF-IDF,position}\}} P_{\text{trait}}\left(\text{trait}(\text{terme})|\text{non}\right)$$

L'un des avantages de la classification naïve bayésienne est que chaque distribution est supposée indépendante. L'ajout de nouveaux traits dans la méthode KEA est donc très aisé.

Parmi les variantes de KEA proposées, Frank *et al.* (1999) ajoutent un troisième trait : le nombre de fois que le terme candidat est un terme-clé dans le corpus d'apprentissage. L'ajout de ce trait permet d'améliorer les performances de la version originale de KEA, mais uniquement lorsque la quantité de données d'apprentissage est très importante. Une autre amélioration de KEA, proposée par Turney (2003), tente d'augmenter la cohérence entre les termes candidats les mieux classés. Pour ce faire, une première étape de classification est effectuée avec la méthode originale. Cette première étape permet d'obtenir un premier classement des termes candidats selon leur score de vraisemblance. Ensuite, de nouveaux traits sont ajoutés et une nouvelle étape de classification est lancée. Les nouveaux traits ont pour but d'augmenter le score de vraisemblance des termes candidats ayant un fort lien sémantique avec certains des termes les mieux classés après la première étape. Enfin, Nguyen et Kan (2007) proposent l'ajout des

9. Sarkar *et al.* (2012) proposent une étude comparative de l'usage des arbres de décision, de la classification naïve bayésienne et des réseaux de neurones pour l'extraction automatique de termes-clés.

10. Il est important de noter que le score de vraisemblance pour chaque terme candidat permet aussi de les ordonner entre eux.

informations concernant la structure des documents. En effet, certaines sections telles que l'introduction et la conclusion dans les articles scientifiques sont plus susceptibles de contenir des termes-clés qu'une section présentant des résultats expérimentaux, par exemple. Dans leur version modifiée de KEA, ils proposent aussi l'usage de traits linguistiques tels que les parties du discours qui ont prouvées jouer un rôle non-négligeable pour l'extraction de termes-clés (Hulth, 2003).

En même temps que KEA (Witten *et al.*, 1999), Turney (1999) met au point l'algorithme génétique GenEx. GenEx est constitué de deux composants. Le premier composant, le géniteur, sert à apprendre des paramètres lors de la phase d'apprentissage. Ces paramètres sont utilisés par le second composant, l'extracteur, pour donner un score d'importance à chaque terme candidat. Plus les paramètres sont optimaux, meilleure est la classification des termes. Pour ce faire, les paramètres sont représentés sous la forme de bits qui constituent une population d'individus que le géniteur fait évoluer jusqu'à obtenir un état stable correspondant aux paramètres optimaux.

Dans son article présentant GenEx, Turney (1999) discute une autre méthode pour l'extraction de termes-clés. Cette méthode utilise de nombreux traits qui servent à entraîner 50 arbres de décision C4.5 (technique de *Random Forest*). Dans un arbre de décision, chaque branche représente un test sur l'un des traits d'un terme candidat. Les tests permettent un routage du terme candidat vers la feuille de l'arbre qui détermine sa classe. Grâce à la technique de *Random Forest*, soit l'usage de plusieurs arbres entraînés sur un échantillon différent du corpus d'apprentissage, l'extraction automatique de termes-clés est réduite à un vote de chaque arbre pour chaque terme candidat. Cela permet un classement des termes candidats en fonction de leur nombre de votes positifs. Les termes-clés extraits correspondent aux termes candidats les mieux classés.

La même année que les travaux de Hulth (2003) sur le bien fondé d'utiliser des traits linguistiques pour l'extraction automatique de termes-clés, Sujian *et al.* (2003) proposent une méthode utilisant un modèle d'entropie maximale (cf. équation 8) dont l'un des traits repose sur les parties du discours des mots qui composent les termes candidats. Un modèle de maximum d'entropie consiste à trouver parmi plusieurs distributions, une pour chaque trait, laquelle a la plus forte entropie. La distribution ayant la plus forte entropie est par définition celle qui contient le moins d'informations, ce qui la rend de ce fait moins arbitraire pour l'extraction des termes-clés.

$$\text{Score}(\text{terme}) = \frac{P(\text{oui}|\text{terme})}{P(\text{non}|\text{terme})} \tag{8}$$

$$P(\text{classe}|\text{terme}) = \frac{\exp\left(\sum_{\text{trait}} \alpha_{\text{trait}} \times \text{trait}(\text{terme}, \text{classe})\right)}{\sum_{c \in \{\text{oui},\text{non}\}} \exp\left(\sum_{\text{trait}} \alpha_{\text{trait}} \times \text{trait}(\text{terme}, c)\right)}$$

Le paramètre α_{trait} définit l'importance du trait auquel il est associé.

Les Séparateurs à Large Marge sont aussi des classifieurs utilisés par les méthodes d'extraction automatique de termes-clés. Ils exploitent divers traits afin de projeter des exemples et des contres-exemples sur un plan, puis ils cherchent l'hyperplan qui les sépare. Cet hyperplan sert ensuite dans l'analyse de nouvelles données. Dans le contexte de l'extraction de termes-clés, les exemples sont les termes-clés et les contres-exemples sont les termes candidats qui ne sont

pas des termes-clés. Ce mode de fonctionnement des SVM est utilisé par Zhang *et al.* (2006), mais un autre type de SVM est plus largement utilisé dans les méthodes supervisées d'extraction de termes-clés. Il s'agit de SVM qui utilisent de multiples marges représentant des rangs. Ces classifieurs permettent donc d'ordonnancer les termes-clés lors de leur extraction (Herbrich *et al.*, 1999; Joachims, 2006; Jiang *et al.*, 2009). La méthode KeyWE de Eichler et Neumann (2010) utilise ce type de SVM avec le trait TF-IDF ainsi qu'un trait booléen ayant la valeur vraie si le terme candidat apparaît dans un titre d'un article Wikipedia (un terme candidat apparaissant dans le titre d'un article de Wikipedia a une plus forte probabilité d'être un terme-clé). L'ordonnancement des termes candidats par le SVM permet ensuite de contrôler le nombre de termes-clés à extraire (choix des k termes candidats les mieux classés).

Tout comme Turney (1999), Ercan et Cicekli (2007) utilisent eux aussi une forêt d'arbres C4.5 dans leur méthode d'extraction de termes-clés. Ils utilisent des traits classiques et leur contribution se situe au niveau de l'utilisation d'un trait calculé à partir de chaînes lexicales. Une chaîne lexicale lie les mots d'un document selon certaines relations telles que la synonymie, l'hyponymie ou la méronymie. Ces relations permettent de calculer un score qui sert de trait. Cette approche est intéressante, mais du fait de limitations des chaînes lexicales actuellement disponibles elle présente l'inconvénient de ne retourner que des mots (aucun multi-mot). Cependant, l'usage d'une forêt d'arbre C4.5 permet un classement des mots à partir de leur nombre de votes positifs. Il est donc envisageable de déduire les termes-clés à partir de la liste ordonnée et pondérée des mots clés (voir les méthodes non-supervisées à bases de graphe – section 2.1).

Une autre méthode pour l'extraction automatique de termes-clés consiste à utiliser un perceptron multi-couches (Sarkar *et al.*, 2010). Un perceptron multi-couches est un réseau de neurones constitué d'au moins trois couches, chaque couche étant composée de neurones. Dans les deux couches extrêmes les neurones représentent respectivement les entrées et les sorties. Les couches centrales sont des couches cachées qui permettent d'acheminer les valeurs des entrées vers les sorties, où de nouvelles valeurs sont obtenues grâce à la pondération des transitions d'un neurone d'une couche vers un neurone de la couche suivante. Les entrées correspondent aux traits d'un terme candidat (ici TF-IDF, la position, la taille, etc.) et les sorties représentent les classes qu'il peut prendre (terme-clé ou non terme-clé). La valeur obtenue pour chaque sortie (classe) permet d'obtenir une probabilité pour que le terme candidat analysé soit un terme-clé ou non. Dans leur méthode, Sarkar *et al.* (2010) utilisent cette probabilité pour ordonner les termes candidats afin de mieux contrôler le nombre de termes-clés à extraire.

Dans leurs travaux, Liu *et al.* (2011) proposent une méthode d'extraction de termes-clés basée sur un modèle génératif. Leur méthode est très différente de celle de Witten *et al.* (1999) puisqu'ils décident d'utiliser une approche de traduction automatique. L'usage original de cette approche est justifié par le fait qu'un ensemble de termes-clés doit décrire de manière synthétique le document. Leur hypothèse est donc qu'un ensemble de termes-clés est une traduction d'un document dans un autre langage. Le modèle est appris à partir de paires de traductions dont l'un des termes est issu des titres ou des résumés des documents du corpus d'apprentissage et dont l'autre terme est issu des corps de ces mêmes documents. Les titres et les résumés sont utilisés comme langage synthétique et les corps des documents comme le langage naturel de ceux-ci.

3 Conclusion et perspectives

L'extraction automatique de termes-clés est une tâche importante qui permet la valorisation d'un document (représentation synthétique, mise en évidence des points clés dans le document, etc.) et qui facilite l'accès aux documents pertinents pour une requête utilisateur (indexation pour la recherche d'information).

Les méthodes existantes pour la tâche d'extraction automatique de termes-clés sont soit supervisées, soit non-supervisées. Les méthodes non-supervisées sont des méthodes émergentes ayant la particularité de s'abstraire de la spécificité des données traitées. Cette abstraction s'explique par des approches basées sur des constatations à propos de ce qu'est un terme-clé au sens général : importance sémantique, degré d'information, structure syntaxique, etc. Contrairement aux méthodes non-supervisées, les méthodes supervisées n'utilisent pas de propriétés définies à partir des traits statistiques et linguistiques, mais elles utilisent des modèles de décision appris à partir de ces traits, calculés sur les termes-clés d'un corpus d'apprentissage. L'usage d'un corpus d'apprentissage implique que les modèles appris soient spécifiques au domaine disciplinaire et à la langue de celui-ci. Cette spécificité peut s'avérer avantageuse lorsque le domaine et la langue que représente le corpus sont les mêmes pour les documents qui sont ensuite analysés, mais si tel n'est pas le cas les résultats de l'extraction peuvent en pâtir.

De futurs travaux peuvent se focaliser sur une hybridation des méthodes non-supervisées et supervisées. Dans un premier temps, il peut être intéressant de tenter d'améliorer les méthodes à base de graphe existantes. En effet, le graphe possède plusieurs points de variabilité sur lesquels il est possible d'agir pour affiner l'extraction : la granularité des nœuds, le type de relations permettant la création des arcs ou encore le facteur d'atténuation λ utilisé dans le calcul du score des nœuds. La granularité peut être étendue à des groupes de phrasèmes similaires (des variantes dont le sens est sensiblement le même). Cette nouvelle granularité peut impliquer la définition d'une nouvelle relation pour la création des arcs entre les nœuds. Enfin, des traits peuvent être appris, pondérés grâce à de l'apprentissage préalable, puis utilisés avec le facteur $(1 - \lambda)$ dans le calcul du score pour chaque nœud (voir la modification du score dans TopicalPageRank (Liu *et al.*, 2010)). Il est possible que ce dernier point demande de modifier la formule du score PageRank afin d'utiliser le score de recommandation et de nouveaux traits de manière cohérente (sans que la valeur d'un trait ne puisse annihiler le score de recommandation).

Remerciements

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

Références

BARKER, K. et CORNACCHIA, N. (2000). Using Noun Phrase Heads to Extract Document Keyphrases. *In Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence : Advances in Artificial Intelligence*.

BOUDIN, F. et MORIN, E. (2013). Keyphrase Extraction for N-best Reranking in Multi-Sentence Compression. *In Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

BRIN, S. et PAGE, L. (1998). The Anatomy of a Large-Scale hypertextual Web Search Engine. *In Proceedings of the 7th International Conference on World Wide Web*.

CLAVEAU, V. (2012). Vectorisation, Okapi et Calcul de Similarité pour le TAL : pour Oublier Enfin le TF-IDF. *In Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN*.

DING, Z., ZHANG, Q. et HUANG, X. (2011). Keyphrase Extraction from Online News Using Binary Integer Programming. *In Proceedings of 5th International Joint Conference on Natural Language Processing*.

EICHLER, K. et NEUMANN, G. (2010). DFKI KeyWE : Ranking Keyphrases Extracted from Scientific Articles. *In Proceedings of the 5th International Workshop on Semantic Evaluation*.

ERCAN, G. et CICEKLI, I. (2007). Using Lexical Chains for Keyword Extraction.

FRANK, E., PAYNTER, G., WITTEN, I., GUTWIN, C. et NEVILL-MANNING, C. (1999). Domain-Specific Keyphrase Extraction.

HASAN, K. et 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*.

HERBRICH, R., GRAEPEL, T. et OBERMAYER, K. (1999). Support Vector Learning for Ordinal Regression. *In Artificial Neural Networks, 1999*.

HULTH, A. (2003). Improved Automatic Keyword Extraction Given More Linguistic Knowledge. *In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*.

JIANG, X., HU, Y. et LI, H. (2009). A Ranking Approach to Keyphrase Extraction. *In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*.

JOACHIMS, T. (2006). Training Linear SVMs in Linear Time. *In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*.

JONES, K. (1972). A Statistical Interpretation of Term Specificity and its Application in Retrieval.

JONES, S. et STAVELEY, M. (1999). Phrasier : a System for Interactive Document Retrieval Using Keyphrases. *In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*.

KIM, S. N., MEDELYAN, O., KAN, M. et BALDWIN, T. (2010). Semeval-2010 Task 5 : Automatic Keyphrase Extraction from Scientific Articles. *In Proceedings of the 5th International Workshop on Semantic Evaluation*.

LIANG, W., HUANG, C., LI, M. et LU, B. (2009). Extracting Keyphrases from Chinese News Articles Using Textrank and Query Log Knowledge. *In Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*.

LITVAK, M. et LAST, M. (2008). Graph-Based Keyword Extraction for Single-Document Summarization. *In Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization*.

LIU, Z., CHEN, X., ZHENG, Y. et SUN, M. (2011). Automatic Keyphrase Extraction by Bridging Vocabulary Gap. *In Proceedings of the 15th Conference on Computational Natural Language Learning*.

LIU, Z., HUANG, W., ZHENG, Y. et SUN, M. (2010). Automatic Keyphrase Extraction via Topic Decomposition. *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

LIU, Z., LI, P., ZHENG, Y. et SUN, M. (2009). Clustering to Find Exemplar Terms for Keyphrase Extraction. *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 1*.

MATSUO, Y. et ISHIZUKA, M. (2004). Keyword Extraction from a Single Document Using Word Co-occurrence Statistical Information.

MEDELYAN, O. et WITTEN, I. (2008). Domain-Independent Automatic Keyphrase Indexing with Small Training Sets.

MIHALCEA, R. et TARAU, P. (2004). Textrank : Bringing Order Into Texts. *In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*.

NGUYEN, T. et KAN, M. (2007). Keyphrase Extraction in Scientific Publications. *In Proceedings of the 10th international conference on Asian digital libraries : looking back 10 years and forging new frontiers*.

NOBATA, C., COTTER, P., OKAZAKI, N., REA, B., SASAKI, Y., TSURUOKA, Y., TSUJII, J. et ANANIADOU, S. (2008). Kleio : a Knowledge-enriched Information Retrieval System for Biology. *In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*.

PAROUBEK, P., ZWEIGENBAUM, P., FOREST, D. et GROUIN, C. (2012). Indexation Libre et Contrôlée d'Articles Scientifiques Présentation et Résultats du Défi Fouille de Textes DEFT2012.

PAUKKERI, M. et HONKELA, T. (2010). Likey : Unsupervised Language-Independent Keyphrase Extraction. *In Proceedings of the 5th International Workshop on Semantic Evaluation*.

ROBERTSON, S. E., WALKER, S., BEAULIEU, M. et WILLETT, P. (1999). Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive Track.

SARKAR, K., NASIPURI, M. et GHOSE, S. (2010). A New Approach to Keyphrase Extraction Using Neural Networks.

SARKAR, K., NASIPURI, M. et GHOSE, S. (2012). Machine Learning Based Keyphrase Extraction : Comparing Decision Trees, Naïve Bayes, and Artificial Neural Networks.

SUJIAN, L., HOUFENG, W., SHIWEN, Y. et CHENGSHENG, X. (2003). News-Oriented Keyword Indexing with Maximum Entropy Principle.

TOMOKIYO, T. et HURST, M. (2003). A Language Model Approach to Keyphrase Extraction. *In Proceedings of the ACL 2003 workshop on Multiword expressions : analysis, acquisition and treatment-Volume 18*.

TURNEY, P. (1999). Learning Algorithms for Keyphrase Extraction.

TURNEY, P. (2003). Coherent Keyphrase Extraction via Web Mining.

WAN, X. et XIAO, J. (2008a). Collabrank : Towards a Collaborative Approach to Single-Document Keyphrase Extraction. *In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*.

WAN, X. et XIAO, J. (2008b). Single Document Keyphrase Extraction Using Neighborhood Knowledge. *In Proceedings of Association for the Advancement of Artificial Intelligence*.

WAN, X., YANG, J. et XIAO, J. (2007). Towards an Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction. *In Annual Meeting-association For Computational Linguistics*.

WITTEN, I., PAYNTER, G., FRANK, E., GUTWIN, C. et NEVILL-MANNING, C. (1999). KEA : Practical Automatic Keyphrase Extraction. *In Proceedings of the 4th ACM conference on Digital libraries*.

ZHANG, K., XU, H., TANG, J. et LI, J. (2006). Keyword Extraction Using Support Vector Machine.