Intégration des plongements de mots dans les méthodes, supervisées et non supervisées, d'extraction automatique de mots clés

Josiane Mothe, Michel Rajoelina, Faneva Ramiandrisoa, Hary Razakasoa

Intégration des plongements de mots dans les méthodes, supervisées et non supervisées, d'extraction automatique de mots clés

Josiane MOTHE (*,***), **Michel RAJOELINA** (**), **Faneva RAMIANDRISOA** (*,**), **Hary RAZAKASOA** (**)
Josiane.Mothe@irit.fr, mrajoel@yahoo.fr, Faneva.Ramiandrisoa@irit.fr, razakasoaharymahefa@gmail.com

(*) IRIT, Université de Toulouse, France,
(**) Université d'Antananarivo, Madagascar,
(***) ESPE, UT2J, France.

Mots clefs:
Extraction automatique de mots clés, plongement de mots, méthodes supervisées, méthodes non supervisées
Keywords:
Automatic keywords extraction, word embedding, management, supervised methods, unsupervised methods
Palabras clave:
Extracción automática de palabras clave, palabra incrustada, métodos supervisados, métodos no supervisados

Résumé

Le plongement de mots a été utilisé avec succès dans diverses applications dans les domaines de traitement de langue et de recherche d'information. Ce papier vise à analyser l'impact de l'intégration des plongements de mots dans les méthodes supervisées et non supervisées d'extraction automatique de mots clés. Les méthodes à base de graphe pour les méthodes non supervisées et les méthodes à base d'ensemble d'arbres de décision pour les méthodes supervisées sont très utilisées et étudiées compte tenu de leurs performances ; nous nous concentrons donc sur celles-ci. Nous avons considéré Word2Vec [24], une méthode de plongement de mots et nous avons évalué l'impact de l'intégration du plongement de mots sur deux jeux de données qui sont des références dans la littérature. Nous avons montré qu'il n'y a pas de différence significative dans les résultats quand nous intégrons le plongement de mots dans les méthodes non supervisées à base de graphe. Pour les méthodes supervisées à base d'ensemble d'arbres de décision, l'intégration du plongement de mots améliore significativement les résultats pour trois des quatre méthodes que nous avons testées. Cet article est une extension des articles [25, 26] qui ne s'intéressaient qu'aux méthodes non supervisées.

1 Introduction

Les *mots clés* sont des mots ou multi-mots qui permettent de caractériser le contenu d'un document. Dans le cas des publications scientifiques qui nous intéresse, les mots clés permettant d'avoir une vue d'ensemble rapide du sujet du document. Ils peuvent également être utilisés comme une entrée de recherche, en recherche d'information, traitement du langage naturel et exploration de texte. L'affectation des mots clés à un document peut être manuelle ou automatique.

Pour les articles scientifiques, il y a habituellement trois types de termes associés: (a) les mots clés libres qui sont fournis par les auteurs, (b) les mots clés choisis par les documentalistes ou par les auteurs à partir d'un thésaurus ou d'une ressource ontologique proposé par l'éditeur et (c) des index de texte libre qui sont des mots ou des groupes de mots extraits automatiquement du contenu du document, comme le font les moteurs de recherche pour les documents du Web.

Ces trois types de mots clés jouent un rôle important pour les systèmes automatiques, par exemple pour la recherche d'information, les études scientométriques, les revues de la littérature ou la classification de documents. Tout en ayant des objectifs communs, ces trois types de mots clés sont très différents. Les mots clés des auteurs sont très appropriés pour représenter le contenu du document, puisque les auteurs sont des spécialistes du sujet sur lequel ils écrivent; mais d'un autre côté, le choix des termes est très subjectif et différents mots clés peuvent être utilisés pour différents textes sur le même sujet. Les mots clés extraits à partir de ressources ontologiques ou de thésaurus n'ont pas cet inconvénient puisque les mots clés sont choisis dans une liste limitée de termes et selon des critères définis (exhaustivité, spécificité). D'un autre côté, les ressources ontologiques sont difficiles à mettre à jour et donc leur contenu peut ne pas refléter l'état actuel d'un domaine ou ne pas inclure des termes spécifiques qui seraient utiles pour décrire précisément le contenu du document. Enfin, les mots clés extraits automatiquement correspondent aux termes d'indexation. Ce dernier processus d'extraction est basé sur des mots extraits du contenu du document. Selon les techniques utilisées, certains prétraitements peuvent rendre les termes choisis non compréhensibles par les humains (par exemple, lorsque l'indexation est à base de radicaux choisis après la racinisation des mots). D'autres techniques au contraire préservent l'intelligibilité des résultats ; c'est le cas de l'extraction de groupes de mots automatique, sujet de cet article.

Dans cet article nous nous focalisons sur les méthodes automatiques d'extraction de groupes de mots, compréhensibles et lisibles par les humains. Ces méthodes visent à extraire automatiquement un nombre limité de mots ou de groupes de mots à partir des textes. Afin d'évaluer les résultats, les mots clés fournis par les auteurs sont généralement considérés comme des vérités terrain.

Dans la littérature, plusieurs méthodes ont été proposées pour extraire automatiquement des mots clés, supervisées ou non. L'avantage des méthodes non supervisées par rapport aux méthodes supervisées est qu'elles n'ont pas besoin d'un ensemble d'apprentissage; par conséquent, elles sont moins sensibles aux changements de sujet et donc plus adaptables. Les méthodes supervisées quant à elles fournissent de meilleurs résultats lorsqu'elles sont utilisées sur des jeux de données pour lesquels l'apprentissage est représentatif.

Cet article vise plus spécifiquement à étudier l'intégration du plongement de mots dans les méthodes d'extraction de mots clés, supervisées ou non, et ses impacts. En effet, le plongement de mots est une méthode récente qui a été utilisée dans plusieurs applications, mais il n'existe pas à notre connaissance d'études relatives à leur intégration dans l'extraction automatique de mots clés.

Dans ce papier, nous nous focalisons sur les méthodes à base de graphe pour les méthodes non supervisées et sur les méthodes à base d'ensemble d'arbres de décision pour les méthodes supervisées. Nous nous intéressons à ces méthodes car elles sont très utilisées et étudiées compte tenu de leur efficacité. Nous montrons d'abord comment le plongement de mots peut être intégré dans les modèles d'extraction de mots clés; nous détaillons cette intégration en considérant les méthodes basées sur les graphes puis les méthodes à base d'ensemble d'arbres de décision. Nous comparons ensuite les résultats obtenus en considérant plusieurs collections et en étudiant différents paramètres.

Le reste de l'article est organisé de la manière suivante: la section 2 présente les méthodes d'extraction automatique de mots clés dans la littérature. La section 3 rapporte comment nous intégrons le plongement de mots dans les méthodes d'extraction de mots clés. La section 4 présente le cadre d'évaluation et les résultats. Enfin la section 5 conclut ce papier.

2 Méthodes de l'état de l'art

Une méthode d'extraction de mots clés se divise généralement en deux étapes: (1) extraire une liste de mots ou groupes de mots qui servent de mots clés candidats et (2) déterminer parmi ceux-ci lesquels sont effectivement des mots clés. Dans la littérature, différentes méthodes d'extraction de mots clés candidats ont été proposées ; elles utilisent généralement des règles heuristiques comme l'utilisation des N-gram [33, 29, 4], des syntagmes non récursifs (parties nominales) [14] ou des patrons grammaticaux prédéfinis [13, 32]. Les méthodes d'extraction de mots clés peuvent être classées en deux catégories : non supervisées et supervisées.

Les *méthodes non supervisées* sont formulées comme des problèmes d'ordonnancement : les mots clés candidats sont ordonnés selon leurs scores d'importance et les N premiers sont considérés comme mots clés. Selon [12], les méthodes non supervisées peuvent être à leur tour classées en quatre catégories : les méthodes à base de graphe [23, 22, 11, 4, 25], les méthodes de regroupement thématique [11, 20], les méthodes d'apprentissage simultané [35, 32] et les méthodes basées sur un modèle de langue [29].

Les *méthodes supervisées* quant à elles ont d'abord été formulées comme des problèmes de classification : chaque mot clé candidat est classé soit comme un mot clé, soit comme un mot non clé. L'objectif est d'entraîner un classifieur sur des documents annotés manuellement avec des mots clés. La création du corpus d'entrainement se fait comme suit : pour chaque document, les mots clés candidats qui font partie des annotations manuelles sont classés comme des exemples positifs et ceux n'appartenant pas à des annotations manuelles sont classés comme des contre-exemples ou exemples négatifs. Différents algorithmes d'apprentissage peuvent être utilisés pour entrainer ce classifieur, notamment les classifications naïve bayésienne [7, 34], les arbres de décision [30, 31], le bagging [14], le boosting [15], l'entropie maximale [18], le perceptron multicouche [21] et les machines à vecteur support [16]. Cependant, comme ces méthodes considèrent le problème comme une classification binaire, ces méthodes ne permettent pas d'ordonner les mots clés candidats ou d'identifier quels mots clés candidats sont meilleurs que d'autres. Certaines méthodes supervisées ordonnent les mots clés candidats au lieu de les catégoriser. C'est le cas de l'approche de [16]. Les auteurs proposent une approche par paire de mots-clés dans laquelle l'algorithme doit ordonner deux mots clés candidats. Cette approche introduit donc la compétition entre les mots clés candidats et a montré des résultats qui surpassent significativement ceux de KEA [7, 34], un ensemble de méthodes supervisées qui adoptent l'approche de classification supervisée traditionnelle [17].

3 Intégration du plongement de mots dans les méthodes d'extraction de mots clés

3.1 Plongement de mots

Avant le plongement de mots (*word embedding* en anglais), la représentation en sac de mots (simples) était la plus couramment utilisée dans les applications traitant des textes. Cependant, la représentation en sac de mots ne capture pas les relations entre les mots. Le plongement de mots est une solution qui permet de pallier ce problème. Le plongement de mots est basé sur l'hypothèse que les mots utilisés dans les mêmes contextes tendent à avoir des significations similaires et donc prennent en compte des relations sémantiques entre les mots. Les méthodes de plongements de mots peuvent être catégorisées en deux types [1] : basées sur le comptage et basées sur les réseaux de neurones.

Ces deux types diffèrent lors de la construction des vecteurs de mots ainsi que par le contexte qu'elles prennent en compte. Les méthodes basées sur le comptage utilisent les documents comme contexte et capturent la similarité sémantique entre documents alors que celles basées sur les réseaux de neurones utilisent les mots voisins comme cotexte pour détecter la similarité mot à mot. Les approches basées sur le comptage tendent à être utilisées pour la modélisation de sujets car elles capturent très bien la relation sémantique, alors que les approches basées sur les réseaux de neurones sont plus efficaces pour obtenir la similarité entre les mots.

Dans ce travail, nous nous intéressons à l'utilisation des approches basées sur les réseaux de neurones car elles permettent de capturer de meilleures relations mot à mot. Plus précisément, nous avons considéré Word2Vec [24], une méthode de plongement de mots qui a prouvé son efficacité dans plusieurs tâches de traitement

automatique de langue. Dans nos expériences, nous avons utilisé le modèle pré-entrainé de Google[1] qui contient des vecteurs de dimensions 300 pour 3 millions de mots et de groupe de mots. Avec ce modèle pré-entrainé, nous pouvons représenter un mot simple avec un vecteur de dimension égale à 300. Dans la section suivante, nous présentons la façon dont Word2Vec est inclus dans les méthodes d'extraction de mots clés.

3.2 Principe d'intégration du plongement de mots dans les méthodes d'extraction de mots clés

3.2.1 Méthodes non supervisées

Pour les méthodes non supervisées nous nous intéressons aux méthodes à base de graphe car elles sont très étudiées, utilisées et fournissent de bons résultats. Les principales étapes de ces méthodes à base de graphe sont :

- **Prétraitements** : Stanford POS Tagger est utilisé pour étiqueter les mots dans le document ; seuls les noms et adjectifs sont retenus pour la suite.
- **Construction du graphe de mots** : un graphe de mots est construit à partir des mots résultant de la première étape. Chaque nœud représente un mot et deux nœuds sont reliés si les mots qu'ils représentent cooccurrent dans une fenêtre fixe dans le document. Nous avons étudié trois méthodes de pondération durant nos expérimentations qui différent par la façon dont le poids de l'arête entre deux nœuds (mots simple) est calculé :
  (i) *Cooccurrence* : le poids de l'arête est le nombre de cooccurrences des deux mots (nœuds) dans une fenêtre de cooccurrence de dix mots comme dans les travaux de Boudin [4].
  (ii) *Cooccurrence avec Word2Vec* : le poids de l'arête est obtenu par le produit du nombre de cooccurrences par la similarité cosinus des deux vecteurs représentant les deux mots (nœuds). L'idée est de renforcer le lien sémantique entre deux mots par le nombre de fois qu'ils cooccurrent dans le document.
  (iii) *Word2Vec seulement* : le poids de l'arête est la valeur de la similarité cosinus des deux vecteurs représentant les deux mots (nœuds). L'intuition derrière cette approche est de s'appuyer entièrement sur la représentation Word2Vec pour quantifier la relation entre deux nœuds.
- **Ordonnancement des nœuds** : Des algorithmes d'ordonnancement de nœuds sont appliqués sur le graphe de mots ainsi construit. Avec cette étape, chaque nœud (mot) se voit attribuer un score. Durant nos expérimentations nous avons utilisé et comparé les mesures de centralité suivantes pour ordonner les nœuds : TextRank [23], Hits [19], Vecteur propre [3], Proximité [2], Intermédiarité [8] and la centralité de degré [4]. Nous avons aussi modifié la méthode RAKE [27] qui utilise une formule basée sur la centralité de degré pour ordonner les nœuds. Pour RAKE, une différence par rapport aux autres approches est la façon dont la cooccurrence est calculée. Dans les autres méthodes, deux nœuds sont reliés s'ils cooccurrent dans une fenêtre de cooccurrence ; dans RAKE, deux nœuds sont reliés s'ils cooccurrent dans les mots clés candidats. Cela permet à RAKE de faire abstraction d'une fenêtre glissante de taille arbitraire.
- **Construction et ordonnancement des mots clés candidats** : Les mots clés candidats sont les séquences adjacentes des mots, restreints aux noms et adjectifs, dans le document. Le score d'un mot clé candidat est la somme normalisée des scores de chaque mot le composant.
- **Sélection des mots clés** : Les mots clés candidats ayant les meilleurs scores sont considérés comme des mots clés.

3.2.2 Méthodes supervisées

Pour les méthodes supervisées, nous avons optés pour les méthodes à base d'ensemble d'arbres de décision parce qu'elles sont les plus utilisées et les plus étudiées compte tenu de leurs performances. Elles présentent aussi d'autres avantages tels que la possibilité d'interpréter les résultats et la possibilité de sélectionner les variables les plus importantes. Par ailleurs, les algorithmes sont très rapides pour la classification de nouveaux cas. Les méthodes utilisant les arbres de décision utilisent généralement des représentations en sacs de mots (c'est le plus simple) et ne tiennent pas compte de la relation sémantique des mots. Comme pour les

[1] https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/view

approches non supervisées, nous nous sommes intéressés à intégrer les relations sémantiques entre les mots par une représentation vectorielle des mots de type Word2vec. Les principales étapes des méthodes utilisant les arbres de décision sont les suivantes :

- **Apprentissage du modèle** : l'extraction automatique de mots clés avec des méthodes supervisées nécessite un ensemble de données connues, appelé ensemble de données d'apprentissage ou d'entraînement pour créer un modèle de prédiction. Dans cet ensemble de données, chaque document est annoté par un ensemble d'exemples positifs et un ensemble d'exemples négatifs (contre-exemples). L'ensemble d'exemples positifs est l'ensemble des mots servant à former les mots clés assignés par l'auteur du document alors que l'ensemble d'exemples négatifs est l'ensemble des mots servant à former les mots clés candidats qui ont les plus mauvais score d'après la méthode TextRank, c'est à dire les mots clés candidats qui ne sont pas considérés comme mots clés par la méthode TextRank. Nous avons choisi TextRank car il s'agit d'une méthode de référence dans les méthodes non supervisées. Ces exemples (mots simples) sont fusionnés pour former une seule liste d'exemples $L = \{(e_1,y_1), (e_2,y_2), \ldots, (e_n,y_n)\}$ de taille n telle que e_i est un mot et $y_i \in \{0,1\}$ est son étiquette. y_i vaut 1 si e_i appartient à l'ensemble des exemples positifs et 0 sinon. Chaque document dans l'ensemble d'entrainement est ainsi représenté par une liste d'exemples.

Pour construire le modèle, nous transformons d'abord notre ensemble de données d'entrainement et pour cela nous utilisons deux types de représentations :

(i) *Sacs de mots* : chaque exemple (mot simple) e_i est représenté par un vecteur $X_i = (0,0,\ldots,x_k=1,\ldots,0)$ de taille v^2 et k est le rang de l'exemple e_i. Un document est représenté comme dans le tableau 1.

(ii) *Représentation vectorielle avec Word2Vec* : chaque exemple (mot simple) e_i est représenté par un vecteur $Xi = (X_i^1, X_i^2, \ldots, X_i^{300})$ avec Word2Vec. Un document est représenté comme dans le tableau 2 tel que chaque ligne est la représentation vectorielle d'un exemple avec Word2Vec :

Tableau 1 : Représentation en sac de mots des exemples (mots simples) constituant les mots clés des auteurs et les mots non clés extraits avec TextRank. v est la taille du vocabulaire (nombre de mots distinct). y est l'étiquette de l'exemple.

Rang / Exemple	1	…	k		v	y
e_1 (si rang = 1)	1	…	0	…	0	y_1
…	…	…	…	…	…	…
e_i	0	…	1	…	0	y_i
…	…	…	…	…	…	…
e_n (si rang = v)	0	…	0	…	1	y_n

Tableau 2: Représentation vectorielle avec Word2Vec des exemples (mots simples) constituant les mots clés des auteurs et les mots non clés extraits avec TextRank. y est l'étiquette de l'exemple.

Exemple	1	2	…	300	y
e_1	X_1^1	X_1^2	…	X_1^{300}	y_1
…	…	…	…	…	
e_i	X_i^1	X_i^2	…	X_i^{300}	y_i

[2] La taille du vocabulaire.

…	…	…	…	…	
e_n	X_n^1	X_n^2	…	X_n^{300}	y_n

Durant l'entrainement, chaque document dans le jeu de données d'entrainement est représenté soit selon une représentation du type de celle du tableau 1 pour créer les modèles utilisant les sacs de mots, soit selon celle correspondant au tableau 2 pour créer les modèles utilisant le plongement de mots. Les arbres de décision sont élaborés à partir de ces tableaux dans le jeu d'entrainement. Durant nos expérimentations, nous avons utilisé et comparé les algorithmes suivants : ExtraTree ou arbre extrêmement aléatoire [10], Adaptative Boosting (AdaBoost) [9], Random Forests ou forêts aléatoires [5] et le Bagging [6].

- **Construction et ordonnancement des mots clés candidats :** les mots clés candidats sont les séquences adjacentes des mots restreints aux noms et adjectifs. Chaque mot constituant les mots candidats sont transformés soit en représentation en sac de mots, soit en représentation vectorielle avec Word2Vec et ensuite la représentation est donnée comme entrée au modèle pré-entrainé. Pour chaque mot donné en entrée, le modèle fournit en sortie la probabilité pour qu'il soit classé dans la classe mot clé (y = 1). Le score d'un mot clé est la somme normalisée des probabilités associées aux mots le composant.
- **Sélection des mots clés :** les mots clés candidats ayant les meilleurs scores sont considérés comme des mots clés.

4 Comparaison des méthodes

4.1 Jeu de données

Dans nos expérimentations, nous avons utilisé les jeux de données INSPEC [14] et SEMEVAL [17]. Ce sont les jeux de données les communément utilisés pour évaluer les méthodes d'extraction de mots clés dans la littérature. De plus, ils sont différents l'un de l'autre, ce qui est très important pour évaluer les méthodes d'extraction automatique de mots clés afin de bien appréhender leurs forces et faiblesses selon Hassan [13].

- **INSPEC** [14] est composé de 2 000 résumés d'articles de journaux de 1998 à 2002. Chaque document est composé d'un titre et d'un résumé. Ce corpus est donc un ensemble de documents courts. Il est composé de données d'entrainement (1 000 documents), de validation (500 documents) et de test (500 documents).
- **SEMEVAL** [17] est composé 244 articles scientifiques de la bibliothèque numérique ACM. Chaque document est composé du contenu intégral de l'article. Ce corpus est donc un ensemble de documents longs. Cette collection est aussi divisée en données d'entrainement (144 documents), de validation (40 documents) et de test (100 documents). Il faut noter que l'ensemble de validation est un sous ensemble des données d'entrainement, c'est-à-dire que les 40 documents des données de validation font parties des 144 documents des données d'entrainement. Ce corpus a été utilisé pour la tâche SemEval-2010[3] tâche 5: Extraction automatique de mots clés à partir d'articles scientifiques.

Pour évaluer les méthodes que nous avons implémentées, nous utilisons les données de test de chaque collection, 500 documents pour INSPEC et 100 documents pour SEMEVAL. Les données d'entrainement et de validation sont utilisées comme corpus d'entrainement des approches supervisées.

[3] http://docs.google.com/Doc?id=ddshp584_46gqkkjng4

4.2 Mesure d'évaluation

La performance de chaque méthode d'extraction automatique de mots clés est évaluée en comparant les mots clés assignés manuellement par les auteurs avec ceux extraits automatiquement pour chaque document à travers les mesures suivantes :

- **Rappel(R)** : défini le nombre de mots clés pertinents retrouvés par rapport au nombre total de mots clés de référence du document (mots clés auteurs).
- **Précision (P)** : défini le nombre de mots clés pertinents retrouvés par rapport au nombre total de mots clés extraits.
- **F1-Mesure** : est la moyenne harmonique du Rappel et de la Précision, pondérés de façon égale.

$$\text{F1-Mesure} = 2 \times \frac{R \times P}{R + P}$$

4.3 Résultats

Durant les évaluations, nous avons considéré les 10 (respectivement 15) premiers mots clés candidats retrouvés automatiquement pour le corpus INSPEC (respectivement SEMEVAL). Notre choix est motivé par le nombre moyen de mots clés assignés par les auteurs pour chaque document : proche de 10 pour INSPEC et de 15 pour SEMEVAL (INSPEC : 9.8 / SEMEVAL : 14.7).

Pour les méthodes non supervisées à base de graphe, nous avons utilisé une fenêtre de cooccurrences de 10 mots car cela fournit les meilleurs résultats avec TextRank selon la littérature [32]. Pour les méthodes supervisées basées sur les arbres de décision, nous avons testé différents nombres d'arbres (50, 100, 200, 300 et 400) et nous avons constaté que les méthodes fournissent de meilleurs résultats pour INPEC avec 200 arbres et 100 arbres pour SEMEVAL. Les résultats présentés dans cette section sont obtenus avec ces configurations.

Le tableau 3 présente les résultats que nous avons obtenus pour les méthodes non supervisées à base de graphe en intégrant ou non Word2Vec. Nous pouvons voir que RAKE donne de meilleurs résultats sur les documents longs (SEMEVAL) et la méthode avec la mesure de centralité proximité fournit les meilleurs résultats pour les documents courts (INSPEC). Nous observons aussi qu'il y a des changements lorsque nous intégrons Word2Vec dans les méthodes, mais les écarts ne sont pas significatifs selon le test de Student avec p-value < 0.05.

Tableau 3: Performances des méthodes non supervisées à base de graphe sur deux jeux de données en considérant (i) Cooccurrence, (ii) Cooccurrence avec Word2Vec, et (iii) Word2Vec seulement. La mesure de centralité Degré n'est pas sensible à l'intégration du plongement de mots et est juste reportée en considérant (i). Les valeurs en gras sont les meilleurs résultats pour chacune des collections.

	Méthodes	INSPEC			SEMEVAL		
		Précision	Rappel	F1-Mesure	Précision	Rappel	F1-Mesure
(i)	Degré	0.31	0.38	0.34	0.10	0.10	0.10
	Vecteur propre	0.30	0.35	0.32	0.08	0.09	0.08
	Proximité	0.33	0.39	0.36	0.05	0.05	0.05
	Hits	0.30	0.35	0.32	0.08	0.09	0.08
	Intermédiarité	0.28	0.34	0.31	0.08	0.08	0.08
	TextRank	0.32	0.38	0.35	0.09	0.09	0.09
	RAKE	0.26	0.31	0.28	**0.18**	**0.12**	**0.14**
(ii)	Vecteur propre	0.30	0.36	0.33	0.07	0.08	0.07

	Proximité	**0.34**	**0.40**	**0.37**	0.03	0.03	0.03
	Hits	0.30	0.36	0.33	0.07	0.08	0.07
	Intermédiarité	0.29	0.34	0.31	0.08	0.09	0.08
	TextRank	0.32	0.38	0.35	0.10	0.10	0.10
	RAKE	0.25	0.29	0.27	0.14	0.14	0.14
(iii)	Vecteur propre	0.30	0.36	0.33	0.09	0.09	0.09
	Proximité	0.34	0.40	0.37	0.02	0.02	0.02
	Hits	0.30	0.36	0.33	0.09	0.09	0.09
	Intermédiarité	0.29	0.34	0.31	0.08	0.08	0.08
	TextRank	0.32	0.38	0.35	0.10	0.10	0.10
	RAKE	0.22	0.26	0.24	0.11	0.11	0.11

Le tableau 4 présente les résultats que nous obtenus pour les méthodes supervisées basées sur les arbres de décision en utilisant la représentation en sac de mots et la représentation vectorielle avec Word2Vec. Nous pouvons constater que la méthode AdaBoost fournit les meilleurs résultats en utilisant le jeu de données INSPEC et la méthode ExtraTree avec le jeu de données SEMEVAL. Sur INSPEC, l'intégration de Word2Vec ne fournit pas de meilleurs résultats par rapport à la représentation en sac de mots, les écarts ne sont pas significatifs selon le test de Student avec p-value égale à 0.05. Avec le jeu de données SEMEVAL, l'utilisation de Word2Vec améliore significativement les résultats pour les méthodes ExtraTree et Random Forests selon le test de Student avec p-value < 0.05. Pour la méthode Bagging, l'écart est significatif avec une p-value < 0.1 alors que pour la méthode AdaBoost, l'écart n'est pas significatif.

Tableau 4: Performances des méthodes supervisées basées sur les arbres de décision sur deux jeux de données en considérant (i) Représentation en sac de mots, (ii) Représentation vectorielle avec Word2Vec. Les valeurs en gras sont les meilleurs résultats pour chacune des collections.

	Méthodes	INSPEC			SEMEVAL		
		Précision	Rappel	F1-Mesure	Précision	Rappel	F1-Mesure
(i)	ExtraTree	0.33	0.38	0.35	0.015	0.017	0.016
	AdaBoost	**0.35**	**0.40**	**0.37**	0.013	0.014	0.013
	Random Forests	0.34	0.40	0.37	0.016	0.017	0.016
	Bagging	0.29	0.34	0.31	0.016	0.017	0.016
(ii)	ExtraTree	0.33	0.38	0.35	**0.032**	**0.034**	**0.033**
	AdaBoost	0.29	0.34	0.31	0.017	0.017	0.017
	Random Forests	0.32	0.37	0.34	0.032	0.033	0.032
	Bagging	0.31	0.36	0.33	0.028	0.029	0.028

Nous constatons à partir des résultats que nous avons obtenus que les résultats des méthodes non supervisées à base de graphe peuvent rivaliser avec les méthodes supervisées basées sur les arbres de décision sur le jeu de données INSPEC et fournissent de meilleurs résultats sur le jeu de données SEMEVAL. Ceci est

peut-être dû au fait qu'il y a très peu de données durant l'entrainement sur SEMEVAL. Nous devons aussi améliorer les prétraitements sur les documents longs comme l'exclusion des symboles par exemple.

5 Conclusions

Dans cet article, nous avons étudié l'impact de l'intégration du plongement de mots, plus précisément Word2Vec, dans les méthodes des méthodes d'extraction automatique de mots clés. Nous avons étudié les méthodes à base de graphe pour les méthodes non supervisées et les méthodes basées sur les arbres de décision pour les méthodes supervisées. Les résultats que nous avons obtenus durant nos expérimentations nous montrent que l'utilisation du plongement de mots améliore significativement les méthodes basées sur les arbres de décisions par rapport à l'utilisation des sacs de mots. Pour les méthodes à base de graphe, l'intégration plongement de mots n'apporte ni d'amélioration ni de diminution significative des résultats par rapport aux méthodes qui ne l'utilisent pas. Nous avons aussi vu que les méthodes à base de graphe peuvent rivaliser avec les méthodes supervisées basées sur les arbres de décision sur le jeu de données INSPEC (composé de résumés d'articles) et les surpassent significativement sur le jeu de données SEMEVAL (composé du contenu intégral d'articles scientifiques).

Comme futurs travaux, nous allons d'abord analyser pourquoi les méthodes basées sur les arbres de décision fournissent de mauvaises performances sur le jeu de données SEMEVAL. Nous voulons aussi entraîner un Word2Vec sur des articles scientifiques au lieu d'utiliser le modèle pré-entrainé sur des articles du web de Google[4]. L'idée derrière est que le modèle serait plus représentatif des mots dans nos corpus qui sont des articles scientifiques.

6 Bibliographie

[1] BARONI M., DINU G., KRUSZEWSKI G., *Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors,* In ACL (1), pages 238–247, 2014.

[2] BAVELAS A., *Communication patterns in task-oriented groups,* J. of the acoustical society of America, 1950.

[3] BONACICH P., *Power and centrality: A family of measures,* Am. j. of sociology, pages 1170–1182, 1987.

[4] BOUDIN F., *A comparison of centrality measures for graph-based keyphrase extraction,* In Int. Joint Conf. on NLP, pages 834–838, 2013.

[5] BREIMAN, L., *Random forests*, Machine learning, *45*(1), 5-32, 2001.

[6] BREIMAN, L., *Bagging predictors*, Machine learning, *24*(2), 123-140, 1996.

[7] FRANK E., PAYNTER G. W., WITTEN I. H., GUTWIN C., NEVILL-MANNING C. G., *Domain-specific keyphrase extraction,* In Proceedings of 16th Int. Joint Conf. on Artificial Intelligence, pages 668–673, 1999.

[8] FREEMAN L. C., *A set of measures of centrality based on betweenness,* Sociometry, pages 35–41, 1977.

[9] FREUND Y., SCHAPIRE R. E., *Experiments with a new boosting algorithm,* Icml. Vol. 96. 1996.

[10] GEURTS P., ERNST D., WEHENKEL L., *Extremely randomized trees,* Machine learning, *63*(1), 3-42, 2006.

[11] GRINEVA M., GRINEV M., LIZORKIN D., *Extracting key terms from noisy and multitheme documents*, In Int. Conf. on WWW, pages 661–670, 2009.

[12] HASAN K. S., NG. V., *Automatic keyphrase extraction: A survey of the state of the art*, In ACL (1), pages 1262–1273, 2014.

[13] HASAN K. S., NG. V., *Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art,* In Int. Conf. on Computational Linguistics, pages 365–373. ACL, 2010.

[4] https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/view

[14] HULTH A., *Improved automatic keyword extraction given more linguistic knowledge,* In Proceedings of the 2003 Conf. on Empirical methods in NLP, pages 216–223, ACL, 2003.

[15] HULTH A., KARLGREN J., JONSSON A., BOSTRÖM H., ASKER L., *Automatic keyword extraction using domain knowledge,* In Proceedings of the 2nd Int. Conf. on Computational Linguistics and Intelligent Text Processing, pages 472–482, 2001.

[16] JIANG X., HU Y., LI H., *A ranking approach to keyphrase extraction,* In Proceedings of the 32nd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pages 756–757, 2009.

[17] KIM S. N., MEDELYAN O., KAN M.-Y., BALDWIN T., *Automatic keyphrase extraction from scientific articles,* Language resources and evaluation, 47(3):723–742, 2013.

[18] KIM S. N., KAN M.-Y., *Re-examining automatic keyphrase extraction approaches in scientific articles,* In Proceedings of the ACL-IJCNLP Workshop on Multiword Expressions, pages 9–16, 2009.

[19] KLEINBERG J. M., *Authoritative sources in a hyperlinked environment,* Journal of the ACM (JACM), 46(5):604–632, 1999.

[20] LIU Z., HUANG W., ZHENG Y., SUN M., *Automatic keyphrase extraction via topic decomposition*, In Proceedings of the 2010 Conf. on empirical methods in NLP, pages 366–376, ACL, 2010.

[21] LOPEZ P., ROMARY L., *HUMB: Automatic key term extraction from scientific articles in GROBID*, In Proceedings of the 5th Int. Workshop on Semantic Evaluation, pages 248–251, 2010.

[22] MATSUO Y., ISHIZUKA M., *Keyword extraction from a single document using word co-occurrence statistical information,* Int. Journal on Artificial Intelligence Tools, 13(01):157–169, 2004.

[23] MIHALCEA R., TARAU P., *Textrank: Bringing order into texts,* ACL, 2004.

[24] MIKOLOV T., DEAN J., *Distributed representations of words and phrases and their compositionality,* Advances in Neural Info. Proc. systems, 2013.

[25] MOTHE J., RAMIANDRISOA F., *Extraction automatique de termes-clés : Comparaison de méthodes non supervisées*, RJCRI CORIA, 2016.

[26] MOTHE J., RAMIANDRISOA F, RASOLOMANANA M. *Automatic Keyphrase Extraction using Graph-based Methods. ACM Symposium on Applied Computing (SAC 2018)*, 2018

[27] ROSE S., ENGEL D., CRAMER N., COWLEY W., *Automatic keyword extraction from individual documents*, Text Mining: Applications and Theory, pages 1–20, 2010.

[28] SONG M., SONG I. Y., HU X., *KPSpotter: a flexible information gain-based keyphrase extraction system,* In Proceedings of the 5th ACM Int. workshop on Web information and data management (pp. 50-53), ACM, 2003.

[29] TOMOKIYO T., HURST M., *A language model approach to keyphrase extraction*, In Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18, pages 33–40, ACL, 2003.

[30] TURNEY, P. D., *Learning algorithms for keyphrase extraction*, Information Retrieval, 2(4) :303–336, 2000.

[31] TURNEY, P. D., *Learning to extract keyphrases from text,* National Research Council Canada, Institute for Information Technology, Technical Report ERB-1057, 1999.

[32] WAN X., XIAO J., *Single document keyphrase extraction using neighborhood knowledge*, In AAAI, volume 8, pages 855–860, 2008.

[33] WAN X., YANG J., XIAO J., *Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction*, In ACL, volume 7, pages 552–559, 2007.

[34] WITTEN I. H., PAYNTER G. W., FRANK E., GUTWIN C., AND NEVILL-MANNING C. G., *Kea : Practical automatic keyphrase extraction,* In Proceedings of the fourth ACM Conf. on Digital libraries, pages 254–255, ACM, 1999.

[35] ZHA H., *Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering*, In Proceedings of the 25th annual Int. ACM SIGIR Conf. on Research and development in information retrieval, pages 113–120, ACM, 2002.