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

Ygor Gallina

▶ **To cite this version:**

Ygor Gallina. État de l'art des méthodes d'apprentissage profond pour l'extraction automatique de termes-clés. 21e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL), Jul 2019, Toulouse, France. hal-02395693

HAL Id: hal-02395693
https://hal.science/hal-02395693

Submitted on 5 Dec 2019

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'apprentissage profond pour l'extraction automatique de termes-clés

Ygor Gallina
LS2N, Université de Nantes, France
ygor.gallina@univ-nantes.fr

RÉSUMÉ

Les termes-clés facilitent la recherche de documents dans de larges collections de données. Le coût d'annotation de document en termes-clés très élevé, c'est pourquoi les chercheurs s'intéressent à cette problématique. Dans cet article nous présentons un état de l'art sur l'extraction automatique de termes-clés en nous intéressant particulièrement aux modèles d'apprentissage profond. En effet, la récente publication d'un demi-million de documents annotés à permis le développement de modèles neuronaux profonds.

ABSTRACT

State of the art of deep learning methods for automatic keyphrase extraction

The use of keyphrases benefits document retrieval in large collection of data. The high cost of annotating documents with keyphrases leads researchers to address this issue. In this article, we outline a state of the art regarding the automatic extraction of keyphrases while focusing on deep learning models. Indeed, the recent availability of half a million of annotated documents made the development of deep neural models possible.

MOTS-CLÉS : extraction de termes-clés ; apprentissage profond ; état de l'art.

KEYWORDS: keyphrase extraction ; deep learning ; state of the art.

1 Introduction

Les termes-clés, également appelés mots-clés, sont des mots ou expressions polylexicales qui représentent les concepts importants d'un document (Evans & Zhai, 1996). Ils se présentent généralement sous la forme de syntagmes nominaux non récursifs contenant majoritairement des noms communs et adjectifs qualificatifs et relationnels (Daille, 2000). Les termes-clés sont essentiellement utilisés pour indexer les documents et naviguer dans les bibliothèques numériques (Witten *et al.*, 2009). En effet, ils sont une alternative plus synthétique et abstraite à l'indexation plein texte, qui elle, se limite aux seuls mots présents dans le document. Ils sont aussi utilisés pour d'autres tâches du Traitement Automatique de la Langue comme le résumé automatique (Litvak & Last, 2008) et la classification de documents (Hulth & Megyesi, 2006).

Malgré les nombreux avantages que procurent les termes-clés, très peu de documents en sont pourvus. En effet, le coût d'annotation par des indexeurs professionnels (ingénieurs documentalistes) est très élevé. C'est pourquoi la communauté scientifique s'intéresse à l'extraction automatique de termes-clés, notamment au travers de campagnes d'évaluations telles que SemEval (Kim *et al.*, 2010; Augenstein

Espace : la station spatiale chinoise Tiangong-1 est tombée dans le Pacifique

4 avril 2018. – Le lundi 2 avril, les autorités spatiales chinoises ont confirmé le désorbitage du laboratoire spatial Tiangong-1. La majeure partie de la structure s'est consumée lors de la réentrée atmosphérique avant que les restes ne s'écrasent dans l'Océan Pacifique à 8 h 15 heure de Pékin (0 h 15 UTC).

Tiangong-1 («Palais Céleste 1»), était la première station spatiale chinoise. Lancée en 2011, les autorités ont rappelés les derniers astronautes en 2013 avant d'en perdre le contrôle en 2016.

Selon les rapports, la deuxième station spatiale chinoise, Tiangong-2, reste en orbite. La Chine prévoit de lancer une station spatiale permanente en 2022, tandis que les États-Unis songent à se retirer du financement de la Station spatiale internationale à l'horizon 2024.

Termes-clés extraits : Tiangong-1 ; Océan pacifique ; réentrée athmosphérique.

Espace : la station spatiale chinoise Tiangong-1 est tombée dans le Pacifique

4 avril 2018. – Le lundi 2 avril, les autorités spatiales chinoises ont confirmé le désorbitage du laboratoire spatial Tiangong-1. La majeure partie de la structure s'est consumée lors de la réentrée atmosphérique avant que les restes ne s'écrasent dans l'Océan Pacifique à 8 h 15 heure de Pékin (0 h 15 UTC).

Tiangong-1 («Palais Céleste 1»), était la première station spatiale chinoise. Lancée en 2011, les autorités ont rappelés les derniers astronautes en 2013 avant d'en perdre le contrôle en 2016.

Selon les rapports, la deuxième station spatiale chinoise, Tiangong-2, reste en orbite. La Chine prévoit de lancer une station spatiale permanente en 2022, tandis que les États-Unis songent à se retirer du financement de la Station spatiale internationale à l'horizon 2024.

Termes-clés extraits : station spatiale ; Océan pacifique ; *chute* ; Chine.

FIGURE 1 – Exemple d'indexation par termes-clés d'une brève journalistique par deux annotateurs. Les concepts similaires sont surligné avec la même couleur. Les termes-clés extraits absent du document sont écrit en italique.

et al., 2017) [1] ou DEFT (Paroubek *et al.*, 2012; Daille *et al.*, 2017). L'atelier NCA (Gollapalli *et al.*, 2015) (Novel Computationel Approach to Keyphrase Extraction) a été consacré en 2015 à cette tâche. Les conférences traitant des bibliothèques numériques telles que la Joint Conference on Digital Libraries (JCDL) s'y intéressent aussi.

L'extraction automatique de termes-clés consiste a assigner à un document un ensemble de termes-clés extraits à partir du contenu du document ou de ressources externes telles que des thésaurus. Il s'agit d'une tâche complexe de par sa subjectivité. Nous présentons dans la Figure 1 l'annotation d'un article de nouvelles journalistiques en termes-clés par deux lecteurs. Dans cet exemple, bien que les concepts de « station spatiale », d'« océan pacifique » et de « désorbitage » soient capturés par les deux annotateurs, seulement un terme-clé est commun aux deux annotations.

1. La tâche décrite par SemEval-2017 consiste à identifier et classifier des relations entre des processus, matériaux et tâches décrites dans des articles scientifiques.

Dans cet article nous présentons un état de l'art des méthodes d'extraction automatiques de termes-clés en se restreignant aux cadres applicatifs des publications scientifiques et des articles journalistiques. Une attention toute particulière sera portée aux modèles neuronaux qui présentent à l'écriture de cet article les meilleures performances.

Le reste de l'article se présente comme suit. La section 2 présente les modèles d'extraction automatique de termes-clés en commençant par les modèles dits traditionnels puis les modèles neuronaux. La section 3 présente les différents jeux de données et les mesures utilisées pour l'évaluation puis les résultats des modèles que nous avons identifiés comme représentatifs. Enfin nous conclurons et présenterons des perspectives de recherches.

2 Modèles d'extraction automatique de termes-clés

Les travaux de la littérature peuvent être séparés en deux catégories, ceux qui proposent des modèles dits traditionnels, qui reposent sur une chaîne de traitement (c.-à-d. segmentation en mots, étiquetage morpho-syntaxique, analyse en dépendances), et ceux qui exploitent des modèles neuronaux dits de bout-en-bout.

2.1 Modèles traditionnels

Les modèles traditionnels reposent sur un schéma de pondération des unités textuelles présentes dans le document. Ils opèrent le plus souvent en deux étapes, la sélection des termes-clés candidats et leur ordonnancement à partir de traits définis manuellement. Lorsqu'ils s'appuient sur une méthode supervisée, ces derniers considèrent l'extraction automatique de termes-clés comme une tâche de classification binaire (terme-clé ou non).

2.1.1 Sélection des termes-clés candidats

L'étape de sélection des termes-clés candidats consiste à identifier, à l'aide d'heuristiques, les mots ou expressions polylexicales du document possédant les caractéristiques d'un terme-clé. Pour cela, deux heuristiques différentes sont couramment utilisées : l'extraction de n-grammes et l'extraction de syntagmes nominaux.

Les n-grammes sont des séquences de n mots adjacents qui apparaissent dans le document. La nature exhaustive de l'extraction de n-grammes permet de maximiser le nombre de candidats corrects (p. ex. « station spatiale »), mais aussi le nombre de candidats incorrects (p. ex. « spatiale chinoise »). Plusieurs stratégies ont été proposées pour résoudre ce problème comme le filtrage des n-grammes contenant des mots outils (Hulth, 2003) ou ceux dont la fréquence est inférieure à un seuil donné (Medelyan *et al.*, 2009).

Les syntagmes nominaux sont des séquences de plusieurs mots constitués d'une tête, qui est un nom, et de satellites, le plus souvent des adjectifs. Des patrons grammaticaux, qu'ils soient définis manuellement (Bougouin *et al.*, 2013) ou inférés à partir de données annotées (Hulth, 2003), sont utilisés pour extraire les syntagmes nominaux du document. À la différence des n-grammes, les syntagmes nominaux sont naturellement des candidats corrects, et ne nécessitent donc pas de filtrage

supplémentaire. En revanche, ils ne représentent qu'une partie limitée des unités textuelles du document, ce qui engendre des problèmes de silence.

2.1.2 Ordonnancement des termes-clés

Une fois les termes-clés candidats identifiés, il s'agit de les pondérer en fonction de leur importance dans le document. Pour cela, les schémas de pondération issus de la recherche d'information comme le TF-IDF (Salton *et al.*, 1975) peuvent être mis en oeuvre. D'autres modèles, proposés spécifiquement pour l'extraction de termes-clés, offrent cependant de meilleures performances.

Witten *et al.* (1999) ont été les premiers à proposer un modèle dédié à l'extraction automatique de termes-clés. Reposant sur un algorithme de classification naïve bayesienne et seulement deux traits (TF-IDF et la position), leur modèle calcule pour chaque candidat la probabilité qu'il soit un terme-clé. Ce travail précurseur a ouvert la voie à d'autres travaux combinant différents algorithmes de classification et/ou proposant de nouveaux traits (Hulth, 2003; Medelyan *et al.*, 2009; Jiang *et al.*, 2009; Sarkar *et al.*, 2010). Jusqu'à très récemment (Meng *et al.*, 2017), peu d'études portant sur les modèles supervisés ont vues le jour car la quantité de données annotées était alors très limitée. C'est la raison pour laquelle de nombreux chercheurs se sont penchés sur le développement de modèles non supervisés.

TextRank (Mihalcea & Tarau, 2004) est sans doute un des modèles non supervisés les plus populaires. Dans ce modèle, le document est représenté sous la forme d'un graphe sur lequel l'algorithme PageRank (Page *et al.*, 1999) est appliqué pour estimer l'importance des mots. L'idée sous-jacente à TextRank est qu'un mot est important s'il cooccurre avec un grand nombre de mots, et si les mots avec lesquels il cooccurre sont eux aussi importants. Parmi les améliorations proposées dans la littérature pour ce modèle, deux directions de recherche émergent. La première concerne l'utilisation de ressources externes pour améliorer la représentation en graphe du document (Wan & Xiao, 2008a; Liu *et al.*, 2010; Gollapalli & Caragea, 2014). La seconde se concentre sur l'amélioration de l'algorithme d'ordonnancement des noeuds du graphe (Bougouin *et al.*, 2013; Florescu & Caragea, 2017; Boudin, 2018).

Certain travaux se démarquent des autres par leurs approches originales. Les travaux de (Tomokiyo & Hurst, 2003), par exemple, considèrent qu'un terme-clé doit être grammatical et informatif et utilisent des modèles de langues pour ordonner les candidats selon ces deux critères. Liu *et al.* (2009), quand à eux, font l'hypothèse que les concepts importants du document sont véhiculés par plusieurs candidats et se fondent sur des regroupements automatiques pour les pondérer.

Si l'on regarde les travaux précédents, peu d'entre eux exploitent la sémantique des termes-clés, et la plupart se contentent de traits de surfaces. Bennani-Smires *et al.* (2018) proposent un premier effort dans cette direction en utilisant les plongements de mots pour ordonner les termes-clés candidats en fonction de leur proximité sémantique au document.

Les modèles présentés dans cette section souffrent de deux écueils : ils ne capturent ni n'exploitent la sémantique des unités textuelles du document ; et l'architecture en chaîne de traitement, sur laquelle ils s'appuient, propage et aggrave les erreurs commises à chaque étape. Pour pallier à ces écueils, les chercheurs se sont orientés vers le développement de modèles neuronaux profonds, que nous décrivons dans la section suivante.

2.2 Modèles neuronaux profond

Contrairement aux modèles présentés précédemment, les modèles neuronaux reposent sur une architecture de bout-en-bout et infèrent des traits à partir de grandes quantités de données d'apprentissage. Accéder à de grandes quantités de données annotées reste cependant compliqué et a longtemps constitué le frein principal au développement de modèles supervisés. Zhang *et al.* (2016) ont été les premiers à apporter une solution à ce problème à partir de données folksonomiques (issues de l'indexation personnelle). Leur idée est de considérer les mots-dièses (*hashtags*) présents dans les microblogs comme une vérité terrain, pour ensuite entraîner un modèle neuronal d'annotation en séquence. Bien que sortant du cadre de cette étude de part les données utilisées, ce travail montre qu'il est possible d'entraîner des modèles neuronaux profonds et a initié une branche de recherche reposant sur l'acquisition massive de données folksonomiques.

Continuant dans cette direction, Meng *et al.* (2017) ont constitué « Kp20k », un corpus comptant un demi-million de résumés d'articles scientifiques annotés en termes-clés par leurs auteurs dans le but d'entraîner un modèle encodeur-décodeur. Il s'agit, à ce jour, de la plus grande collection disponible de données de ce type.

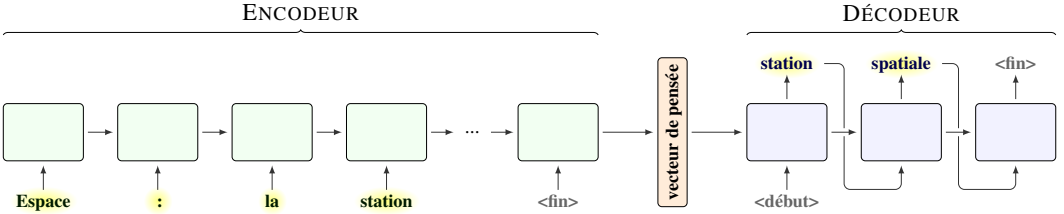

FIGURE 2 – Exemple de modèle *encodeur-décodeur* récurrent appliqué à l'extraction automatique de termes-clés.

Le modèle encodeur-décodeur (Cho *et al.*, 2014; Sutskever *et al.*, 2014) (cf. Figure 2) se décompose en deux parties : un encodeur qui crée un vecteur de pensée (également appelé vecteur latent) à partir de la séquence d'entrée ; et un décodeur qui génère une séquence de sortie. L'encodeur récurrent calcule pour chaque mot un vecteur de pensée conditionné par le vecteur de pensée précédent. Une fois la séquence d'entrée encodée, le décodeur génère séquentiellement les mots en fonction du mot précédemment généré et de son vecteur de pensée.

Ce modèle impose de définir un vocabulaire, certains mots peu fréquents ou non présents dans le corpus d'entraînement ne pourront donc pas être générés. Pour étendre le vocabulaire Gu *et al.* (2016) et See *et al.* (2017) proposent des mécanismes de copie, qui ajoutent la possibilité au réseau de copier un mot à partir de la séquence d'entrée.

La stratégie de décodage utilisée par Meng *et al.* (2017) permet de générer une séquence de mots, donc un seul terme-clé. Pour générer un ensemble de termes-clés, un algorithme de recherche en faisceau (OW & MORTON, 1988) est utilisé. Le décodage est alors effectué à partir des n mots les plus probables au lieu d'un seul. Les termes-clés générés par la recherche en faisceau ne sont pas liés les uns aux autres, ils sont donc souvent redondants. Pour limiter ce phénomène, Chen *et al.* (2018) génèrent directement une séquence de termes-clés divisés par des marqueurs de séparation, et appliquent un mécanisme de revue qui favorise la diversité des mots dans la séquence produite. Bien que bénéfique pour les performances, l'ajout de ces différents mécanismes augmente significativement le nombre de paramètres et en complexifie l'entraînement.

L'entraînement des modèles encodeur-décodeur nécessite de très grandes quantités de données annotées pour atteindre un niveau de performance satisfaisant (Meng *et al.*, 2017). Collecter des données annotées n'est cependant pas si simple, c'est pourquoi ils se sont intéressés à l'apprentissage semi-supervisé. Ye & Wang (2018) proposent deux moyens d'utiliser des données non annotées pour l'entraînement du modèle encodeur-décodeur. Le premier consiste à appliquer des modèles existants sur des données non annotées pour produire de nouvelles données d'entraînement. Le second consiste à entraîner conjointement le modèle à effectuer deux tâches mutuellement bénéfiques : l'extraction de termes-clés et la génération de titre.

3 Résultats expérimentaux

Dans cette section nous présentons les jeux de données ainsi que les mesures d'évaluation principalement utilisées par la communauté. Dans un second temps nous présentons les résultats expérimentaux des modèles de l'état de l'art et mettons en exergue leurs forces et faiblesses.

3.1 Jeux de données et mesures d'évaluation

Les ensembles de données disponibles dans la littérature sont différenciés par rapport à la nature des documents qui les composent : les articles scientifiques, les notices (résumés et titres d'articles scientifiques) et les nouvelles d'actualités. Nous présentons dans la Table 1 un sous-ensemble de ces jeux de données ventilés selon leur nature.

Nom	Lang.	Ann.	#entr.	#test	#mots	#tc	%abs
CSTR (Witten *et al.*, 1999)	en	A	130	500	11 501	5.4	18.7
SemEval-2010 (Kim *et al.*, 2010)	en	$A \cup L$	144	100	7 961	14.7	19.7
Inspec (Hulth, 2003)	en	I	1 000	500	135	9.8	22.4
TermITH-Eval (Bougouin *et al.*, 2016)	fr	I	-	400	164.7	11.8	59.8
KP20k (Meng *et al.*, 2017)	en	A	530 809	20 000	176	5.3	42.6
DUC-2001 (Wan & Xiao, 2008b)	en	L	-	308	847	8.1	3.7
110-PT-BN-KP (Marujo *et al.*, 2013)	pt	L	100	10	439	27.6	7.5

TABLE 1 – Statistiques des jeux de données utilisés. Les documents sont annotés par les auteurs (A), des lecteurs (L) ou des ingénieurs documentalistes (I). Les colonnes #entr. et #test présentent le nombre de documents dans les corpus d'entraînement et de test. Les colonnes #tc et #mots présentent le nombre moyen de termes-clés et de mots par document. La colonne %abs présente le taux de termes-clés qui n'apparaissent pas dans le document (calculé sur le corpus de test).

La plupart des jeux de données sont en anglais, même si quelques initiatives ont vues le jour dans d'autres langues comme le français ou le portugais. Il est important de noter que ces ensembles sont de taille plutôt réduite et que seul Kp20k dispose de suffisamment de documents pour permettre l'entraînement de modèles neuronaux profonds.

Chaque type d'annotateur annote d'une manière différente. En effet, les ingénieurs documentalistes – professionnels de l'indexation – annotent principalement les notices scientifiques. Leur annotation,

coûteuse, constitue une référence de qualité. Ils réalisent l'extraction de termes-clés en domaine ouvert ou fermé. L'extraction de termes-clés en domaine fermé implique d'assigner un terme-clé choisi parmi une liste de référence, par exemple dans un thésaurus. L'extraction de termes-clés en domaine ouvert, aussi utilisée par les auteurs et lecteurs, fait appel à l'expertise de l'annotateur dans le domaine pour le choix des termes-clés. Ce type d'extraction non contraint peut introduire de la variabilité dans les termes-clés choisis. À la différence des ingénieurs documentalistes dont la finalité de l'annotation est l'indexation, les lecteurs choisissent les termes-clés selon leur expertise du domaine et leur usage de cette annotation. Witten *et al.* (1999) fait l'hypothèse que les auteurs choisissent les termes-clés dans le but de maximiser la visibilité de leur article et constitue une référence biaisée par les sujets de recherche populaires au moment du choix de ces termes-clés.

Les modèles traditionnels, présentés dans la section 2, sont majoritairement extractifs. Ils limitent l'extraction de termes-clés aux seuls mots ou expressions polylexicales présentes dans le document. Le taux de termes-clés de référence absent dans le document permet alors de donner une idée des performances maximales de ces modèles.

Le taux de termes-clés absents diffère en fonction de la nature du document. Les nouvelles d'actualités sont des documents courts, pourtant la très grande majorité des termes-clés annotés sont présents dans le document. Ce phénomène s'explique en partie par l'annotation généralement extractive des lecteurs (Wang *et al.*, 2015). Les notices, elles aussi, sont très courtes (≈ 160 mots) et le taux de termes-clés de référence absents est plus élevé que dans les articles scientifiques, en effet elles en sont des représentations synthétiques.

La disponibilité de différents corpus annotés permet d'effectuer une évaluation automatique des modèles, bien moins coûteuse qu'une évaluation manuelle, en comparant la sortie du modèle avec les termes-clés de référence du corpus.

Il n'y a pas de consensus sur les mesures d'évaluation utilisée, mais les plus usités sont la précision, le rappel et la f-mesure calculés à partir des 5 ou 10 premiers termes-clés. Ces mesures ne prennent pas en compte l'ordre des termes-clés extraits, or les méthodes s'astreignent à les ordonner. Pour préférer les modèles classant les termes-clés corrects en premier, des mesures comme la Bpref (Binary Preference), la MRR (Mean Reciprocal Rank), le NDCG (Normalized Discounted Cumulative Gain) et la MAP (Mean Average Precision) sont utilisés par (Basaldella *et al.*, 2016; Liu *et al.*, 2010; Boudin, 2018; Marujo *et al.*, 2013; Chen *et al.*, 2018).

Les termes-clés extraits sont considérés comme corrects s'ils correspondent exactement à la référence. Ce type de correspondance ne prend pas en compte la variabilité lexicale, c'est pourquoi il est d'usage de comparer les formes racinisées des termes-clés. Cela peut parfois introduire des erreurs (p. ex. la forme racinisée de « empire » et « empirique » est « empir »). Une évaluation manuelle permet de prendre en compte d'autres variantes, telles que les variantes lexicales (p. ex. « apprentissage machine » et « apprentissage automatique »), et de mettre en avant la sous estimation des évaluations automatiques (Bougouin *et al.*, 2016).

3.2 Résultats

Nous présentons dans la Table 2 un sous-ensemble représentatif des résultats des modèles de la littérature. Les résultats des modèles présentés sont ceux reportés dans les articles les décrivant.

Nous observons la grande diversité des mesures et des jeux de données utilisés pour l'évaluation des

Collection \ Modèle(s)	F_1	F_1@5	F_1@10	Bpref	MRR	MAP
CSTR						
(Witten *et al.*, 1999)	-	18.3†	**18.4**†	-	-	-
SemEval-2010						
(Boudin, 2018)	-	12.2	14.5	-	-	11.8
(Chen *et al.*, 2018)	-	32.0‡	**32.0**‡	-	-	-
Inspec						
(Hulth, 2003)	33.9	-	-	-	-	-
(Mihalcea & Tarau, 2004)	36.2	-	-	-	-	-
(Liu *et al.*, 2010)	-	24.2	-	27.4	**58.3**	-
(Boudin, 2018)	-	25.9	30.6	-	-	29.2
Kp20k						
(Meng *et al.*, 2017)	-	**32.8**‡	25.5‡	-	-	-
DUC2001						
(Liu *et al.*, 2010)	-	-	31.2	21.4	**63.8**	-

TABLE 2 – Résultats présentés dans les articles. Les meilleurs score reportés pour chaque jeu de données sont mis en gras. † Calculé à partir de la moyenne des vrai positifs. ‡ Scores calculés à partir des termes-clés présent dans le document.

modèles, ce qui rend très complexe la comparaison des modèles entre eux. Bien que la f-mesure soit la métrique la plus utilisée, le choix du rang ne fait pas consensus bien que les modèles récents la rapportent généralement aux rangs 5 et 10. Il est important de noter que les performances de certains modèles sont calculés sans tenir compte des termes-clés de référence absents du contenu du document. Les mesures tenant compte de l'ordre des termes-clés sont très peu utilisées, bien qu'apportant un moyen de comparaison supplémentaire.

Ce tableau montre aussi que les valeurs de f-mesures obtenues varient entre 10 et 30%, ces scores sont très bas par rapport à d'autres tâches de traitement automatique de la langue. Par exemple, le modèle de la boîte à outils Spacy [2] pour la reconnaissance d'entité nommée obtient une valeur de f-mesures de 85%.

4 Discussion/Perspectives

Dans cet article, nous avons présenté l'état actuel de la recherche pour l'extraction automatique de termes-clés.

Beaucoup de travaux s'intéressent à cette tâche, nous avons présenté les modèles traditionnels reposant sur une chaîne de traitement en deux étapes : la sélection des candidats puis leur ordonnancement. La disponibilité du corpus Kp20K a permis le développement de modèles utilisant l'apprentissage profond pour générer des termes-clés. Ces modèles utilisent une architecture de bout-en-bout sans étape de sélection de candidats, permettant ainsi d'outrepasser les propagations d'erreurs liés aux

2. https://spacy.io/models/en

modèles utilisant des chaînes de traitement.

Malgré cela, nous avons montré que les modèles les plus performants atteignent une valeur de f-mesure de 30%. Ces performances s'expliquent par une évaluation sous-optimale ne prenant pas en compte la variabilité lexicale des termes-clés extrait, l'extractivité des modèles proposés, mais aussi par la subjectivité de la tâche. L'évaluation de cette tâche reste un champ de recherche ouvert. Elle peut être améliorée en prenant en compte les variations lexicales liés à la synonymie et aux relations d'hyperonymies, avec l'utilisation de réseaux lexicaux par exemple.

Nos travaux futurs s'intéresseront au développement des modèles neuronaux profonds, notamment à l'amélioration des modèles génératifs et de leur abstractivité. Mais aussi à l'investigation de moyens alternatifs d'appliquer les modèles neuronaux profonds à l'extraction de termes-clés.

Références

AUGENSTEIN I., DAS M., RIEDEL S., VIKRAMAN L. & MCCALLUM A. (2017). SemEval 2017 Task 10 : ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. In *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*, Vancouver, Canada : Association for Computational Linguistics. arXiv : 1704.02853.

BASALDELLA M., CHIARADIA G. & TASSO C. (2016). Evaluating anaphora and coreference resolution to improve automatic keyphrase extraction. In *In Proceedings of COLING 2016, 26th International Conference on Computational Linguistics*, p.Ĩ1, Osaka, Japan.

BENNANI-SMIRES K., MUSAT C., HOSSMANN A., BAERISWYL M. & JAGGI M. (2018). Simple Unsupervised Keyphrase Extraction using Sentence Embeddings. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, p. 221–229, Brussels, Belgium : Association for Computational Linguistics.

BOUDIN F. (2018). Unsupervised Keyphrase Extraction with Multipartite Graphs. In *Proceedings of NAACL-HLT 2018* : Association for Computational Linguistics. arXiv : 1803.08721.

BOUGOUIN A., BARREAUX S., ROMARY L., BOUDIN F. & DAILLE B. (2016). TermITH-Eval : a French Standard-Based Resource for Keyphrase Extraction Evaluation. In *LREC - Language Resources and Evaluation Conference*, Potoroz, Slovenia.

BOUGOUIN A., BOUDIN F. & DAILLE B. (2013). TopicRank : Graph-Based Topic Ranking for Keyphrase Extraction. In *International Joint Conference on Natural Language Processing (IJCNLP)*, p. 543–551, Nagoya, Japan.

CHEN J., ZHANG X., WU Y., YAN Z. & LI Z. (2018). Keyphrase Generation with Correlation Constraints. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium : Association for Computational Linguistics. arXiv : 1808.07185.

CHO K., VAN MERRIENBOER B., GULCEHRE C., BAHDANAU D., BOUGARES F., SCHWENK H. & BENGIO Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv :1406.1078 [cs, stat]*. arXiv : 1406.1078.

DAILLE B. (2000). Morphological Rule Induction for Terminology Acquistion. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING '00, p. 215–221, Stroudsburg, PA, USA : Association for Computational Linguistics. event-place : Saarbrücken, Germany.

DAILLE B., BARREAUX S., BOUGOUIN A., BOUDIN F., CRAM D. & HAZEM A. (2017). Indexation d'articles scientifiques Présentation et résultats du défi fouille de textes DEFT 2016. *Recherche d'information, document et web sémantique*, **17**(1).

EVANS D. A. & ZHAI C. (1996). Noun-phrase Analysis in Unrestricted Text for Information Retrieval. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, p. 17–24, Stroudsburg, PA, USA : Association for Computational Linguistics. event-place : Santa Cruz, California.

FLORESCU C. & CARAGEA C. (2017). PositionRank : An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1105–1115, Vancouver, Canada : Association for Computational Linguistics.

GOLLAPALLI S. D. & CARAGEA C. (2014). Extracting Keyphrases from Research Papers Using Citation Networks. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

S. D. GOLLAPALLI, C. CARAGEA, X. LI & C. L. GILES, Eds. (2015). *Proceedings of the ACL 2015 Workshop on Novel Computational Approaches to Keyphrase Extraction*. Beijing, China : Association for Computational Linguistics.

GU J., LU Z., LI H. & LI V. O. K. (2016). Incorporating Copying Mechanism in Sequence-to-Sequence Learning.

HULTH A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing* -, volume 10, p. 216–223, Not Known : Association for Computational Linguistics.

HULTH A. & MEGYESI B. B. (2006). A Study on Automatically Extracted Keywords in Text Categorization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, p. 537–544, Stroudsburg, PA, USA : Association for Computational Linguistics. event-place : Sydney, Australia.

JIANG X., HU Y. & 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 - SIGIR '09*, p. 756, Boston, MA, USA : ACM Press.

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*, SemEval '10, p. 21–26, Stroudsburg, PA, USA : Association for Computational Linguistics.

LITVAK M. & LAST M. (2008). Graph-based Keyword Extraction for Single-document Summarization. In *Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization*, MMIES '08, p. 17–24, Stroudsburg, PA, USA : Association for Computational Linguistics. event-place : Manchester, United Kingdom.

LIU Z., HUANG W., ZHENG Y. & SUN M. (2010). Automatic Keyphrase Extraction via Topic Decomposition. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, p. 366–376, Stroudsburg, PA, USA : Association for Computational Linguistics.

LIU Z., LI P., ZHENG Y. & 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 - EMNLP '09*, volume 1, p. 257, Singapore : Association for Computational Linguistics.

MARUJO L., GERSHMAN A., CARBONELL J., FREDERKING R. & NETO J. P. (2013). Supervised Topical Key Phrase Extraction of News Stories using Crowdsourcing, Light Filtering and Co-reference Normalization. *arXiv :1306.4886 [cs]*. arXiv : 1306.4886.

MEDELYAN O., FRANK E. & WITTEN I. H. (2009). Human-competitive Tagging Using Automatic Keyphrase Extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 3 - Volume 3*, EMNLP '09, p. 1318–1327, Stroudsburg, PA, USA : Association for Computational Linguistics.

MENG R., ZHAO S., HAN S., HE D., BRUSILOVSKY P. & CHI Y. (2017). Deep keyphrase generation. In *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, volume 1, p. 582–592.

MIHALCEA R. & TARAU P. (2004). TextRank : Bringing Order into Texts. In *Proceedings of {EMNLP-04}and the 2004 Conference on Empirical Methods in Natural Language Processing*, p.8̃, Barcelona, Spain.

OW P. S. & MORTON T. E. (1988). Filtered beam search in scheduling†. *International Journal of Production Research*, **26**(1), 35–62.

PAGE L., BRIN S., MOTWANI R. & WINOGRAD T. (1999). *The PageRank Citation Ranking : Bringing Order to the Web*.

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. In *DEFT@TALN*.

SALTON G., WONG A. & YANG C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, **18**(11), 613–620.

SARKAR K., NASIPURI M. & GHOSE S. (2010). A New Approach to Keyphrase Extraction Using Neural Networks. *International Journal of Computer Science Issues*, **7**(2). arXiv : 1004.3274.

SEE A., LIU P. J. & MANNING C. D. (2017). Get To The Point : Summarization with Pointer-Generator Networks. *arXiv :1704.04368 [cs]*. arXiv : 1704.04368.

SUTSKEVER I., VINYALS O. & LE Q. V. (2014). Sequence to Sequence Learning with Neural Networks. *arXiv :1409.3215 [cs]*. arXiv : 1409.3215.

TOMOKIYO T. & 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, p. 33–40, Not Known : Association for Computational Linguistics.

WAN X. & XIAO J. (2008a). CollabRank : towards a collaborative approach to single-document keyphrase extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics - COLING '08*, volume 1, p. 969–976, Manchester, United Kingdom : Association for Computational Linguistics.

WAN X. & XIAO J. (2008b). Single Document Keyphrase Extraction Using Neighborhood Knowledge. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, p. 855–860, Chicago, Illinois : AAAI Press.

WANG R., LIU W. & MCDONALD C. (2015). Using Word Embeddings to Enhance Keyword Identification for Scientific Publications. In M. A. SHARAF, M. A. CHEEMA & J. QI, Eds., *Databases Theory and Applications*, Lecture Notes in Computer Science, p. 257–268 : Springer International Publishing.

WITTEN I. H., BAINBRIDGE D. & NICHOLS D. M. (2009). *How to Build a Digital Library, Second Edition*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 2nd edition.

WITTEN I. H., PAYNTER G. W., FRANK E., GUTWIN C. & NEVILL-MANNING C. G. (1999). KEA : practical automatic keyphrase extraction. In *Proceedings of the fourth ACM conference on Digital libraries - DL '99*, p. 254–255, Berkeley, California, United States : ACM Press.

YE H. & WANG L. (2018). Semi-Supervised Learning for Neural Keyphrase Generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. arXiv : 1808.06773.

ZHANG Q., WANG Y., GONG Y. & HUANG X. (2016). Keyphrase Extraction Using Deep Recurrent Neural Networks on Twitter. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 836–845, Austin, Texas : Association for Computational Linguistics.