Note technique

keyBoost - Un méta-modèle par consensus guidé pour l'extraction de mot-clefs

BEKKAR Zakaria
zakaria.bekkar@ens-paris-saclay.fr

ENS PARIS-SACLAY - ENSAE PARIS

Juin 2021

Sommaire

1 Introduction

1.1 Contexte

L'un des nombreux rôles des Directions régionales de l'Environnement, de l'Aménagement et du Logement (DREAL) se structure autour de leur statut particulier d'autorité environnementale. En cette qualité, elles sont notamment amenées à formuler des avis sur tout projet à impact environnementale sur leur territoire de compétence.

Une expérimentation conjointe ménée par l'Ecolab, laboratoire d'innovation en intélligence artificielle (IA) affilié au Ministère de la Transition Ecologique, et la DREAL Brétagne explore les potentialités de l'apport de l'IA dans cette tâche par le bias de la production de *preuves de concept* (*PoC*).

Un des traits saillants du travail quotidien des auditeurs en charge de la formulation de ces avis est l'impératif de traiter une grande masse d'informations de façon à la fois fine et diligente : tout manquements aux délais impartis par le réglément européen en la matière pouvant s'assortir de lourdes sanctions financières. Le bénifice de la mise en place d'outils agissant en support de la rédaction et de l'analyse documentaire est donc double, puisqu'il s'exprime tant du point de vue du gain financier que du point de vue de la charge de travail des auditeurs en leur offrant la possibilité de se concentrer sur ce qui fait la véritable valeur ajoutée de leur expertise.

Ce constat à conduit à l'élaboration d'une *PoC* autour de la notion de *sommaire augmenté*. L'ambition affichée et de pouvoir permettre,par le biais d'un interface utilsateur, une extraction structurée et pertinente de l'information contenues dans les études d'impacts qui forment la base documentaire des avis de l'autorité environnementale. L'une des directions prises pour le *sommaire augmenté* a été la volonté de mettre en place un module d'extraction de mot-clefs opérant sur les nombreuses sections composant les études d'impacts. A mon arrivée à l'Ecolab, j'ai hérité de la responsabilité de concevoir et d'implémenter ce module en parallalèle de l'élaboration d'une *PoC* cette fois-ci sur un sytème de recommandation d'avis.

1.2 Brêve revue de littérature dans le domaine de la *keyword extraction*

L'extraction de mots clés (également appelée détection de mots clés/analyse de mots clés/ keyword extraction) fait référence à l'ensemble des techniques d'analyse de texte utilisées pour extraire automatiquement les mots et expressions les plus importantes d'un texte.Elles permettent de reconnaître les principaux sujets abordés et de résumer le contenu d'un texte de manière succinte.

A ce titre, c'est un sous-domaine du traitement naturel du langage (TLN ou NLP) qui s'est peu à peu consolider autour de deux axes : *Machine Learning* vs Methodes Statistiques vs Théorie des Graphes et Modèles Supervisés vs Modèles Non-Supervisés.

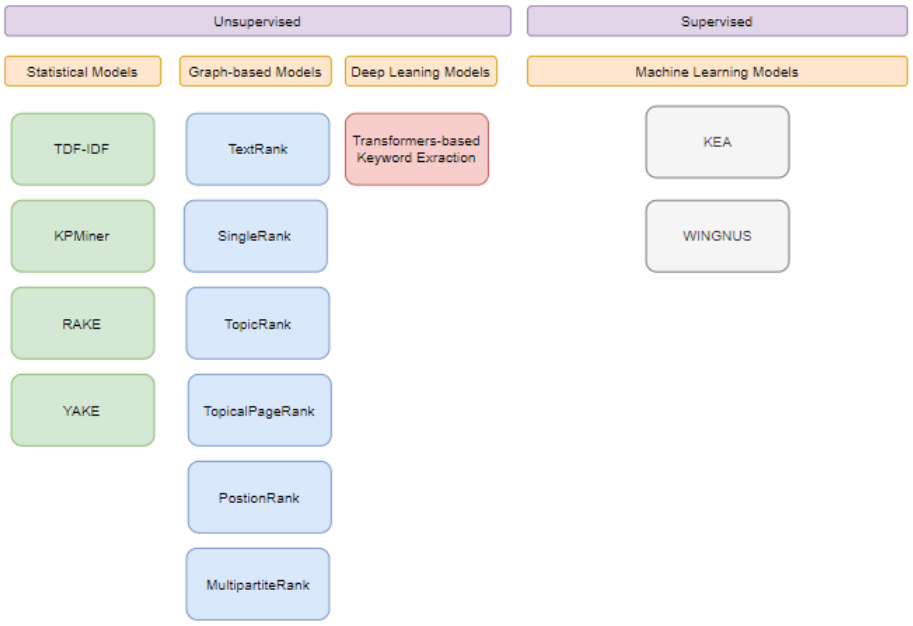

FIGURE 1.1 – Quelques modèles illustrant le paysage du sous-domaine de la keyword extraction

La pléthore de directions dans laquelle la récherche dans ce domaine se fait réflète une réalité à laquelle on peut difficilemment échapper lorsque l'on fait de la keyword extraction : il n'existe pas d'approche univoque surpassant toutes les autres dans tous les scénarios. Les différentes typologies de modèles sont performantes sur différents types de textes.

	Scientific articles						*Paper abstracts*						*News articles*					
	PubMed		**ACM**		**SemEval**		**Inspec**		**WWW**		**KP20k**		**DUC-2001**		**KPCrowd**		**KPTimes**	
Model	F@10	MAP	F@10	MAP	F@10	MAP	F@10	MAP	F@10	MAP	F@10	MAP	F@10	MAP	F@10	MAP	F@10	MAP
FirstPhrases	15.4	14.7	13.6	13.5	13.8	10.5	29.3	27.9	10.2	9.8	13.5	12.6	24.6	22.3	17.1	16.5	9.2	8.4
TextRank	1.8	1.8	2.5	2.4	3.5	2.3	35.8	31.4	8.4	5.6	10.2	7.4	21.5	19.4	7.1	9.5	2.7	2.5
TFxIDF	16.7	16.9	12.1	11.4	17.7	12.7	**36.5**	**34.4**	9.3	10.1	11.6	12.3	23.3	21.6	16.9	15.8	9.6	9.4
PositionRank	4.9	4.6	5.7	4.9	6.8	4.1	34.2	32.2	11.6†	8.4	14.1†	11.2	28.6†	**28.0**†	13.4	12.7	8.5	6.6
MPRank	15.8	15.0	11.6	11.0	14.3	10.6	30.5	29.0	10.8†	10.4	13.6†	13.3†	25.6	24.9†	**18.2**	**17.0**	11.2†	10.1†
EmbedRank	3.7	3.2	2.1	2.1	2.5	2.0	35.6	32.5	10.7†	7.7	12.4	10.0	**29.5**†	27.5†	12.4	12.4	4.0	3.3
Kea	18.6†	18.6†	14.2†	13.3	19.5†	**14.7**†	34.5	33.2	11.0†	10.9†	14.0†	13.8†	26.5†	24.5†	17.3	16.7	11.0†	10.8†
CopyRNN	**24.2**†	**25.4**†	**24.4**†	**26.3**†	**20.3**†	13.8	28.2	26.4	**22.2**†	**24.9**†	**25.4**†	**28.7**†	10.5	7.2	8.4	4.2	**39.3**†	**50.9**†
CorrRNN	20.8†	19.4†	21.1†	20.5†	19.4	10.9	27.9	23.6	19.9†	20.3†	21.8†	22.7	10.5	6.5	7.8	3.2	20.5†	20.3†

FIGURE 1.2 – Evaluation de quelques modèles sur des datasets de benchmarking issus de trois domaines différents (Boudin et al. 2020)

Malgré des efforts de recherches considérables consacrés à ce sujet au fil des ans, le problème de l'extraction de mots-clés pertinents avec une grande précision reste non résolu [bib 20 yake].Tout un ensembles de facteurs sont en cause. Parmi ceux-ci, les difficultés à définir clairement la notion de pertinence ou encore la large diversité linguistique à laquelles les algorithmes sont confrontés (langues, taille, styles de rédaction, domaines etc) sont parmi les défis les plus brûlants.

D'autres obstacles tiennent aux problèmes posé par les mots-clés absents, la restriction de correspondance exacte entre mots-clés labellisés et prédits et le nombre élevé de mots-clés candidats qui peuvent être générés à partir d'un seul texte.Toutes ces questions mettent en évidence les freins au développement d'une solution globale et motivent la nécessité de poursuivre les recherches.

1.3 *keyBoost* : Apports scientifiques et pratiques pour le cas d'usage avec les DREAL

Dans ce contexte, je propose un modèle intitulé *keyBoost* qui est conçu pour palier le problème du choix du modèle de *keyword extraction* en particulier pour les cas d'usages où :

— il n'y a ni données labellisées

— ni recul sur les perfomances potentielles de chaque typologie de modèle

Les données de la DREAL Bretagne, étant de cette nature, offrent un cadre d'application parfait pour cette architecture originale dans la littérature du domaine de la *keyword extraction*

keyBoost se définit en première lieu par son architecture de *méta-modèle par consensus guidé*. Concretement, le modèle se compose d'une selection de modèles considérés *state of the art* pour leur typologie (cf Figure 1.1) et fait en sorte, via un système de consensus qui sera explicité ci-après, de mettre en cohérence l'ensemble des keywords générées par ces modèles pour ne proposer que ceux qui sont globalement les plus susceptibles d'être pertinents. Cette approche validée experimentalement, permet, en l'absence d'a priori sur le meilleur modèle à utilisé, d'avoir une performance plus élevée que la moyenne des sous-modèles utilisés.

2 Fonctionnement de *keyBoost*

2.1 Vue générale du modèle

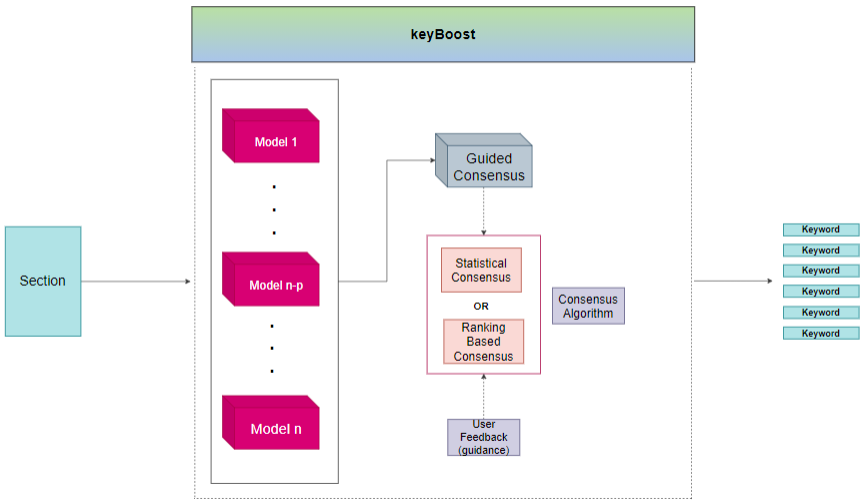

FIGURE 2.1 – Vue d'ensemble de keyBoost

Comme mis en évidence sur le schéma ci-dessus, keyBoost se fonde sur les mots-cléfs génénés par un ensemble de modèle triés sur le volet et en sélectionne les plus pertinents via un *consensus guidé* qui peut s'apparenter intuitivement à un vote.

Le coeur de l'architecture se situe, par voie de conséquence, dans le consensus guidé qui permet d'harmoniser et de juger l'ensemble des sorties des sous-modèles retenues. L'adjectif "guidé" ici à toute son importance puisque ce n'est pas simplement un processus de choix algorithmique deconnecté du contexte de son cas d'usage. Il semblait important de permettre au modèle de prendre ne compte le jugement des utilisateurs sur les mots-cléfs proposés afin de "biaser" consensus des prochaines extraction sur les modèles les plus pertinents.

Dans le cadre du développement de keyBoost, deux approches ont étés explorées parmis les nombreuses possibilités envisageables : une consensus statistique et un consensus basé sur le rang.

2.2 Consensus guidé

2.2.1 *Statistical Consensus*

L'idée du *statistical consensus* est d'exploiter les scores en sorties de tous les modèles shortlistés afin d'avoir une base de comparaison et de sélection.L'obstacle majeur à l'application directe de cette intuition est le fait que ces scores ne sont en réalité pas cohérents les uns avec les autres. Les procédés de génération des scores sont très différents les uns des autres en fonction des architectures utilisées sans parler du sens même que ces quantités peuvent avoir (par exemple la similarité calculée par le modèle de deep learning BERT ou bien encore le score purement statistique en sortie de TF-IDF)

Le *statistical consensus* dépasse cette impossibilité de recouvrer toutes interprétations cardinales globales des scores en tentant de reconstruire statistiquement une interprétabilité ordinale. Cette dernière est largement suffisante pour pouvoir opérer une sélection pertinente des mots-clés.

(Ajouter schéma + equations)
Le procédé retenu est le suivant :

1. Appliquer une transformation à tous les scores pour les faire vivre dans le même espace. (mettre la fonction choisie)
2. Faire en sorte que pour chaque modèle la distribution des scores transformée mettent en évidence la capacité à discriminer les très bon mots clefs de ceux qui sont médiocres
3. Vérifier que la distribution consolidée de tous les scores transformé venant de tous les modèles met en évidence les mêmes propriétés de discrimination
4. Application d'un algorithme de déduplication afin d'éliminer les doublons parmis les keywords qui sont proposés au niveau global
5. Sélection des top k-keywords restant grâce selon le score transformé

2.2.2 *Ranking Based Consensus*

(Ajouter schéma + equations)
Toujours dans cette recherche d'une interprétation ordinale globale la plus fidèle possible de la pertinence de la proposition de keyword de chaque sous-modèle, l'approche *ranking based* se fonde non plus sur les scores mais directement sur les classements.

Selon un nombre k de keywords voulus et n de modèles sous-jacents, la procédure est comme suit :

1. Sélection des $E(\frac{k}{n})$ keywords les plus pertinents selon le score de chaque modèle
2. Application de l'algorithme de déduplication

Une limite de cette algorithme par rapport à l'approche statistique est son caractère rustre. En effet, par construction le *ranking based consensus* ignore les potentielles proposition à partir du $E(\frac{k}{n}) + 1$ rang d'un sous-modèle plus pertinentes que les $E(\frac{k}{n})$ premières d'un autre sous-modèle.

2.2.3 Orientation du consensus via le *feedback* utilisateur

L'idée est d'adjoindre à la procédure de consensus purement algorithmique de l'information issue du domaine via le *feedback utilisateur*,le consensus devient alors *guidé*. (explication de la notion plus globale de l'active Learning, ses propriétes et en quoi keyBoost en tire partie).

Ce feedback se manifeste par un jugement de l'utilisateur final de la *keyword extraction* sur la qualité des propositions de keyBoost

Cette information est ensuite utilisée pour biaiser le consensus en faveur des modèles les plus adaptés au cas d'usage et en l'occurrence ceux jugés les plus pertinents. Les prochains consensus pour le même type de texte mettront en avant ces modèles ce qui se traduira par une surepresentation de leurs keywords.(cas particulier d'active learning. nouveau ?)

(schéma + explicitation de l'algorithme de feedback utilisateur

(prévoir de l'active learning classique via la constitution d'un dataset labelisé ?)

3 Résultats

3.1 Validation de la pertinence de l'architecture keyBoost

3.1.1 Méthodologie

Le point de départ de la méthodologie de validation de l'apport effectif de *keyBoost* est la volonté de mettre en évidence qu'en l'absence d'a priori sur le meilleur modèle à utilisé pour un domaine et un type de texte spécifique, *keyBoost* a une pertinence plus élevée que la moyenne de sous-modèles utilisés.

Cette méthodologie permettrait donc de valider ou pas l'utilité d'un approche ensembliste telle que keyBoost en l'absence de recul sur le meilleur modele ou de données labélisées.

Explication du choix du dataset PubMED + et des deux scénarios de tests sur la configuration de keyBoost.

3.1.2 Résultats

(résultats + discussion et conclusion)

3.2 Sorties typiques pour le cas d'usage avec les DREAL

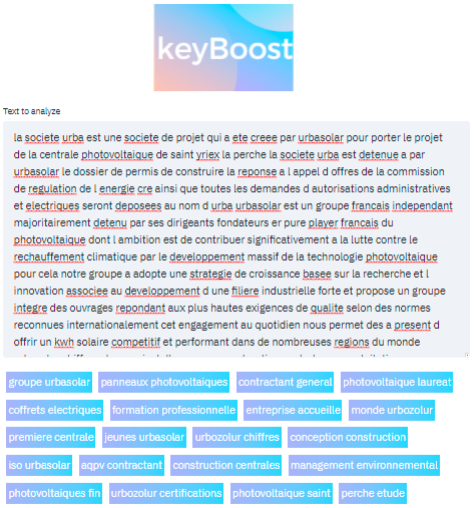

FIGURE 3.1 – Exemple de sortie keyBoost

commentaire

3.3 Le package python *keyBoost*