

Extraction des concepts-clés à partir du fonds Charcot

Approche *PatternRank*

Ljudmila PETKOVIC^{1,2,3}

`prenom.nom@sorbonne-universite.fr`

¹ Sorbonne Université, Faculté des Lettres, UFR Littératures françaises et comparée, ED 3

² Centre d'étude de la langue et des littératures françaises (CELLF), UMR 8599

³ Observatoire des textes, des idées et des corpus (ObTIC)

Atelier ObTIC

DataLab, BNF

Paris, le 24 avril 2024



Plan

1. Contexte de recherche
2. État de l'art
3. *PatternRank* en pratique

1. Contexte de recherche

2. État de l'art

3. *PatternRank* en pratique

Définition de la tâche

Angl. *keyphrases* : « phrases-clés »

- séquences de plusieurs mots (ex. *sclérose latérale amyotrophique*)
- reflètent plus précisément le contexte sémantique du texte
≠ mots-clés, ex. *sclérose*

Extraction

Processus de sélection d'un ensemble de phrases les plus pertinentes à partir d'un texte donné (SCHOPF et al., 2022).

Prédiction

Processus de génération des phrases-clés qui résument parfaitement un document donné (XIE et al., 2023).

1. Contexte de recherche

2. État de l'art

3. *PatternRank* en pratique

Approches classiques (GARAUD, 2022a)

STATISTIQUES

Basées sur les fréquences des mots / groupe de mots et leur cooccurrence.

- TF-IDF – *Term Frequency · Inverse Document Frequency* (SPARCK JONES, 1972)
- RAKE – *Rapid Automatic Keyword Extraction* (ROSE et al., 2010)
- YAKE – *Yet Another Keyword Extractor* (CAMPOS et al., 2020)

À base de GRAPHES

Chaque nœud correspond à un mot / groupe de mots, alors que chaque arc représente la probabilité (ou la fréquence) d'observer ces mots ensemble.

- SingleRank (WAN et XIAO, 2008)
- TextRank (MIHALCEA et TARAU, 2004)
- TopicRank (BOUGOUIN et al., 2013)

Approches sémantiques (GARAUD, 2022b)

Utiliser les « plongements » (angl. *embeddings*) pour lier des mots sémantiquement proches et en extraire ceux qui apportent l'information la plus pertinente.

1 plongements de mots

- réseaux de neurones artificiels « simples »
- représentation vectorielle de l'ensemble des mots d'un vocabulaire
- distance entre ces vecteurs → mots sémantiquement proches
- fastTextRank¹

2 plongements de mots du type BERT / phrases

- réseaux de neurones artificiels « profonds » (angl. *deep learning*)
- basés sur les modèles de langue pré-entraînés
- censés capturer la sémantique du texte
- architecture gourmande en temps de calcul et en volume de données
- ex. : Key2Vec (MAHATA et al., 2018), KeyBERT (GROOTENDORST, 2020)

1. <https://github.com/jeekeim/fasttextrank>

Approches exploitant les modèles de langue – keybert

Librairie Python qui exploite les plongements BERT et la similarité cosinus pour générer les mots/phrases-clés les plus similaires à un document.

- 1 entrée : un document
- 2 tokenisation du document en mots/phrases-clés candidates
- 3 génération des plongements du document et des mots/phrases-clés
- 4 calcul de la similarité cosinus document : mots/phrases-clés

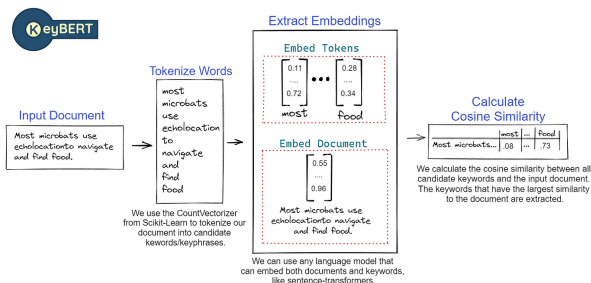


Fig. 1 – Pipeline de la méthode keybert (GROOTENDORST, 2020).

Approches exploitant les modèles de langue – *PatternRank*

KeyBERT + Keyphrase-Vectorizers = ***PatternRank*** (SCHOPF et al., 2022)

- extraction des phrases-clés les plus similaires à un document
- préservation de leur grammaticalité grâce aux motifs POS

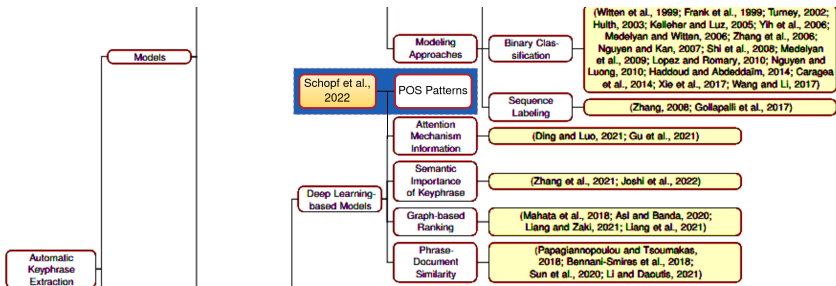


Fig. 2 – Extrait de l'état de l'art sur l'extraction des mots-clés, adapté de XIE et al. (2023)

Fonctionnement de *PatternRank*

- 1 entrée : un seul document texte tokenisé
- 2 étiquetage des tokens avec les balises POS
- 3 sélection des tokens correspondant au modèle POS défini comme phrases-clés candidates
- 4 génération des plongements du document et des phrases-clés candidates par un modèle de langue
- 5 calcul des similarités cosinus entre les plongements du document et des phrases-clés candidates + classement des phrases-clés
- 6 extraction des N phrases-clés les plus représentatives

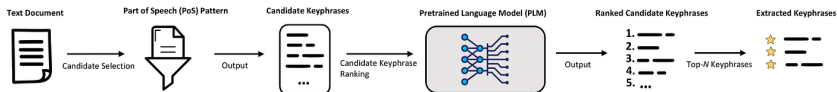


Fig. 3 – Workflow de la méthode *PatternRank*.

1. Contexte de recherche

2. État de l'art

3. *PatternRank* en pratique

Extraction des phrases-clés avec keybert et keyphrase-vectorizers

Ressources en ligne :

- [Lien Google Colab](#)

pré-requis :

- bonne connexion Internet
- mémoire RAM suffisante
- [Dépôt GitHub](#)

Références I



BOUGOUIN, A., F. BOUDIN et B. DAILLE (2013). TopicRank : Graph-Based Topic Ranking for Keyphrase Extraction. In : *International Joint Conference on Natural Language Processing (IJCNLP)*, p. 543-551 ([voir p. 6](#)).



CAMPOS, R., V. MANGARAVITE, A. PASQUALI, A. JORGE, C. NUNES et A. JATOWT (2020). YAKE! Keyword extraction from single documents using multiple local features. In : *Information Sciences* 509, p. 257-289 ([voir p. 6](#)).



GARAUD, D. (22 fév. 2022a). Extraire automatiquement les concepts et mots-clés d'un texte (Part I : Les méthodes dites classiques). Oncrawl. (Visité le 09/04/2024) ([voir p. 6](#)).



GARAUD, D. (22 fév. 2022b). Extraire automatiquement les concepts et mots-clés d'un texte (Part II : approche sémantique). Oncrawl. (Visité le 09/04/2024) ([voir p. 7](#)).



GROOTENDORST, M. (2020). KeyBERT : Minimal keyword extraction with BERT. Version v0.3.0 ([voir pp. 7, 8](#)).



MAHATA, D., J. KURIAKOSE, R. SHAH et R. ZIMMERMANN (2018). Key2Vec : Automatic Ranked Keyphrase Extraction from Scientific Articles using Phrase Embeddings. In : *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, p. 634-639 ([voir p. 7](#)).



MIHALCEA, R. et P. TARAU (2004). TextRank : Bringing Order into Text. In : *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Sous la dir. de D. LIN et D. WU. Barcelona, Spain : Association for Computational Linguistics, p. 404-411 ([voir p. 6](#)).

Références II



ROSE, S., D. ENGEL, N. CRAMER et W. COWLEY (2010). Automatic Keyword Extraction from Individual Documents. In : *Text Mining : Applications and Theory*, p. 1-20 (voir p. 6).



SCHOPF, T., S. KLIMEK et F. MATTHES (2022). PatternRank : Leveraging Pretrained Language Models and Part of Speech for Unsupervised Keyphrase Extraction. In : *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2022) – KDIR*. INSTICC. SciTePress, p. 243-248 (voir pp. 4, 9).



SPARCK JONES, K. (1972). A Statistical Interpretation of Term Specificity and its Application in Retrieval. In : *Journal of documentation* 28.1, p. 11-21 (voir p. 6).



WAN, X. et J. XIAO (août 2008). CollabRank : Towards a Collaborative Approach to Single-Document Keyphrase Extraction. In : *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Sous la dir. de D. SCOTT et H. USZKOREIT. Manchester, UK : Coling 2008 Organizing Committee, p. 969-976 (voir p. 6).



XIE, B., J. SONG, L. SHAO, S. WU, X. WEI, B. YANG, H. LIN, J. XIE et J. SU (2023). From Statistical Methods to Deep Learning, Automatic Keyphrase Prediction : A Survey. In : *Information Processing & Management* 60.4, p. 103382 (voir pp. 4, 9).