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

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 : unigrammes de mot, 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).

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)

Graphes

Chaque nœud = mot / groupe de mots ;
chaque arc = 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)

Plongements de mots

Représentent l'ensemble des mots d'un vocabulaire sous forme de vecteurs. Distance entre ces vecteurs → mots sémantiquement proches.

- `fastTextRank`[1]

Plongements contextuels

Basés sur les modèles de langue pré-entraînés.
Gèrent mieux des cas ambiguës (homographes).

- Key2Vec (Mahata et al., 2018)
- KeyBERT (Grootendorst, 2020)

1. https://github.com/jeekim/fasttextrank

Fonctionnement de la librairie `keybert`

1. entrée : un document
2. tokénisation 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

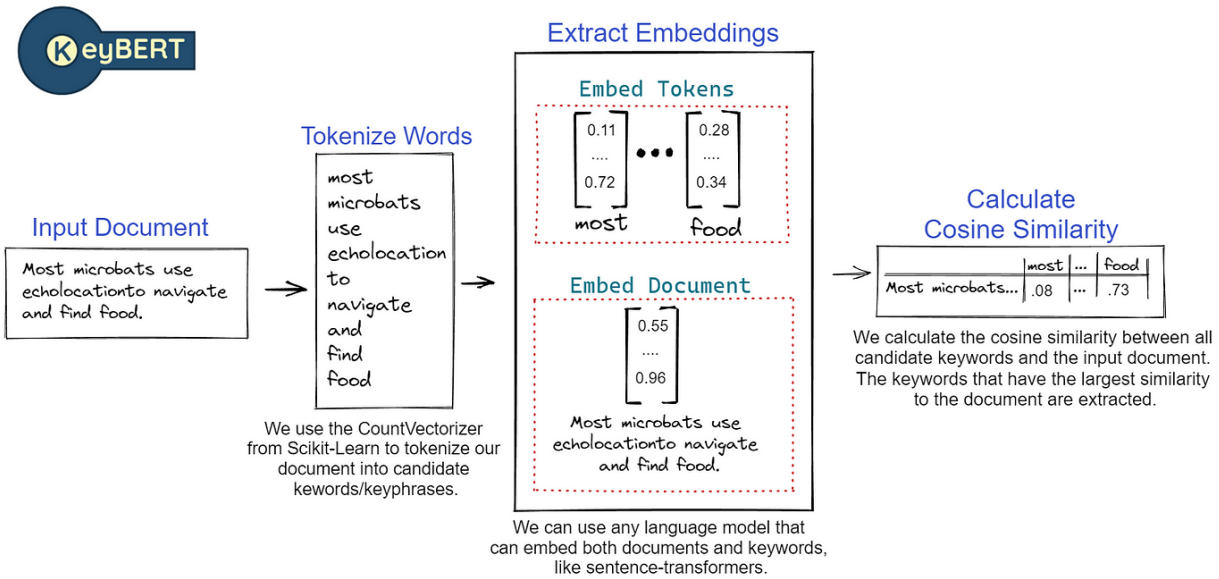

Fig. 1 – *Pipeline* de la méthode `keybert` (GROOTENDORST, 2020).

`keybert` amélioré – *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

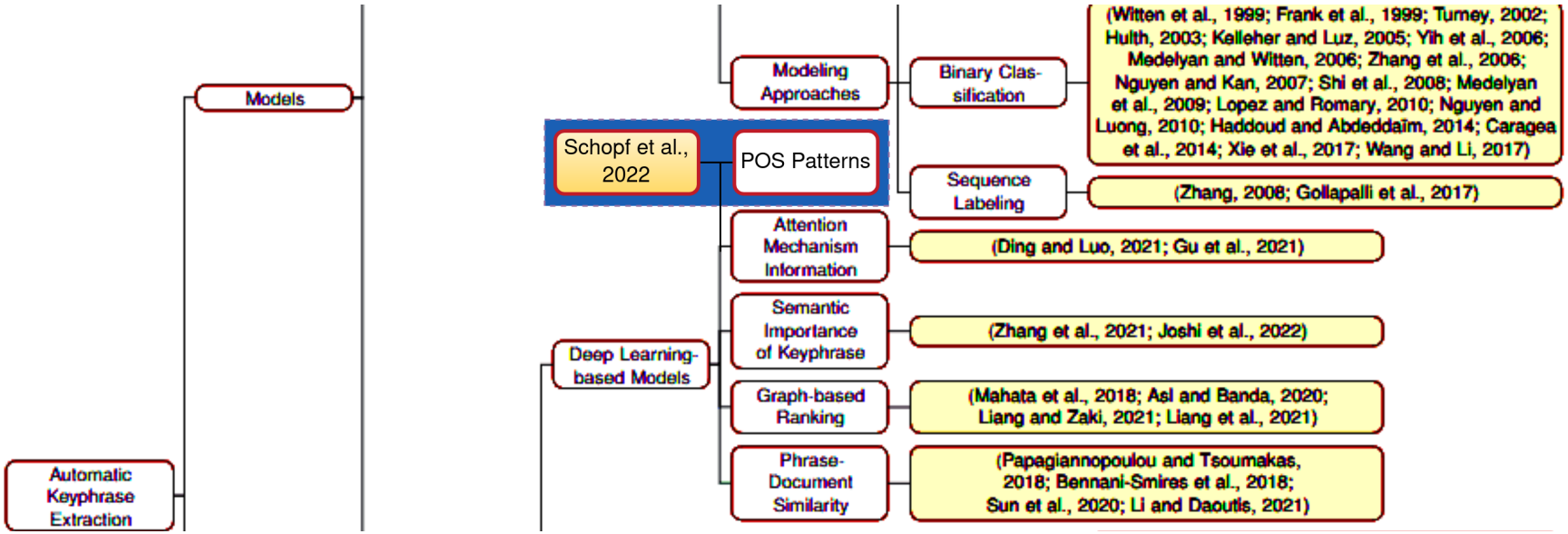

Fig. 2 – Extrait de l'état de l'art sur l'extraction des mots-clés, adapté de XIE et al. (2023)

Fonctionnement de la méthode *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

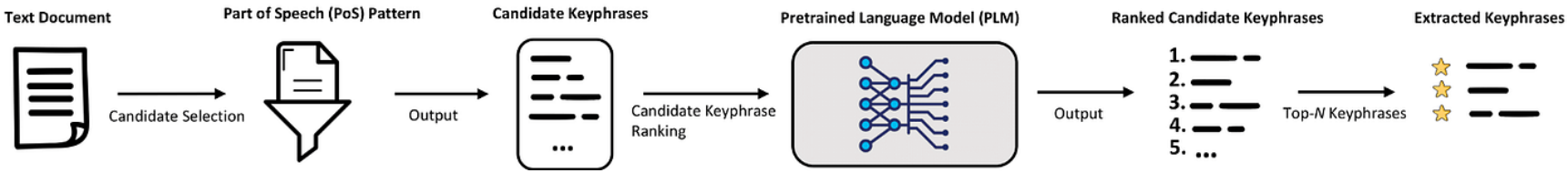

Fig. 3 – *Workflow* de la méthode *PatternRank*.

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*).