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 30 avril 2024

Plan

Librairie `keybert`

- extraction des mots/phrases-clés les plus similaires à un document
- exploitation des plongements BERT

⚠ la longueur des n-grammes à extraire n'est pas inférée en amont

- `keyphrase_ngram_range=(1, 3)` : uni-, bi- ou trigrammes

⚠ la grammaticalité des phrases n'est pas prise en compte

- p. ex. «scientifique les planches»

Librairie `keybert`

- extraction des mots/phrases-clés les plus similaires à un document
- exploitation des plongements BERT

⚠ la longueur des n-grammes à extraire n'est pas inférée en amont

- `keyphrase_ngram_range=(1, 3)` : uni-, bi- ou trigrammes

⚠ la grammaticalité des phrases n'est pas prise en compte

- p. ex. «scientifique les planches»

Librairie `keybert`

- extraction des mots/phrases-clés les plus similaires à un document
- exploitation des plongements BERT

⚠ la longueur des n-grammes à extraire n'est pas inférée en amont

- `keyphrase_ngram_range=(1, 3)` : uni-, bi- ou trigrammes

⚠ la grammaticalité des phrases n'est pas prise en compte

- p. ex. «scientifique les planches»

Librairie `keybert`

- extraction des mots/phrases-clés les plus similaires à un document
- exploitation des plongements BERT

⚠ la longueur des n-grammes à extraire n'est pas inférée en amont
  - `keyphrase_ngram_range=(1, 3)` : uni-, bi- ou trigrammes

⚠ la grammaticalité des phrases n'est pas prise en compte
  - p. ex. «scientifique les planches»

Librairie `keybert`

- extraction des mots/phrases-clés les plus similaires à un document
- exploitation des plongements BERT

⚠ la longueur des n-grammes à extraire n'est pas inférée en amont

- `keyphrase_ngram_range=(1, 3)` : uni-, bi- ou trigrammes

⚠ la grammaticalité des phrases n'est pas prise en compte

- p. ex. « scientifique les planches »

Librairie `keybert`

- extraction des mots/phrases-clés les plus similaires à un document
- exploitation des plongements BERT

⚠ la longueur des n-grammes à extraire n'est pas inférée en amont

- `keyphrase_ngram_range=(1, 3)` : uni-, bi- ou trigrammes

⚠ la grammaticalité des phrases n'est pas prise en compte

- p. ex. « scientifique les planches »

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

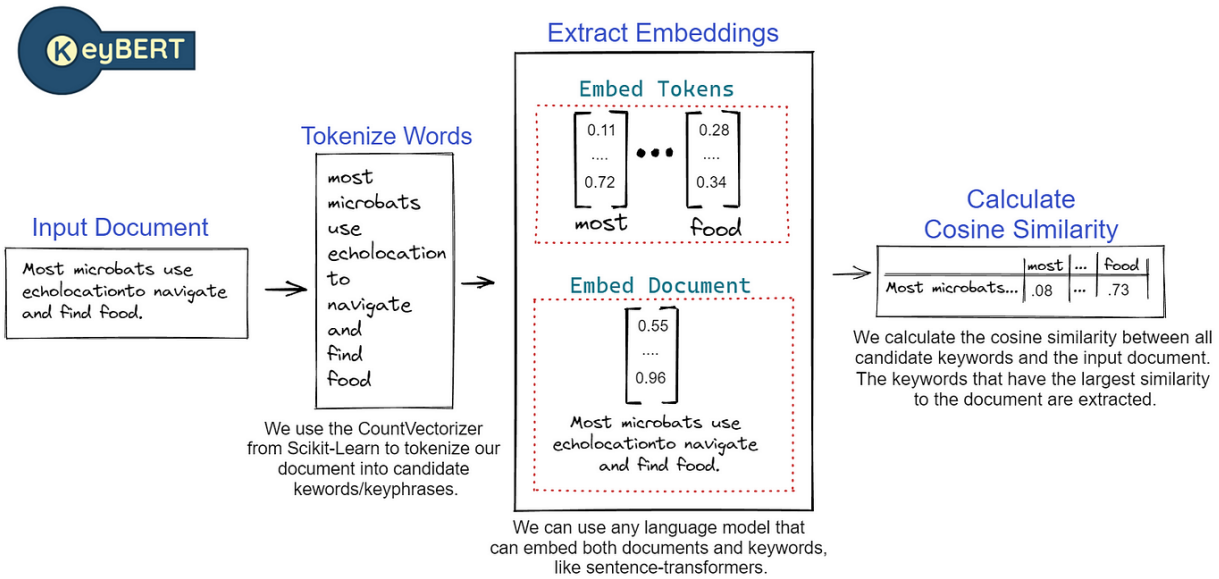

Fig. 1 – *Pipeline* de la méthode `keybert` (GROOTENDORST, 2020).

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

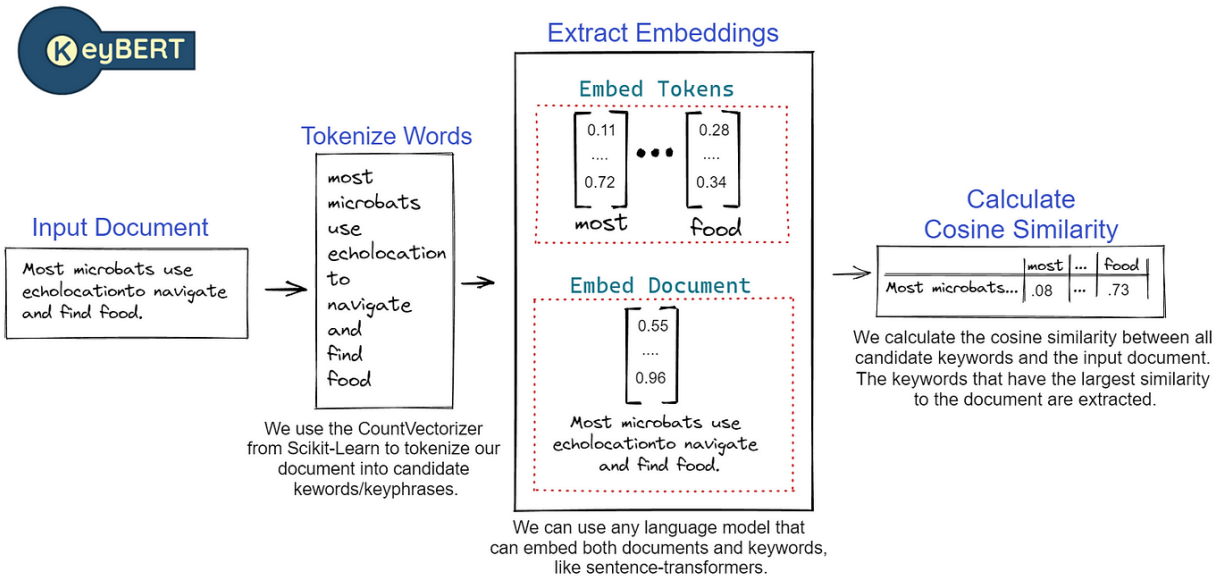

Fig. 1 – *Pipeline* de la méthode `keybert` (GROOTENDORST, 2020).

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

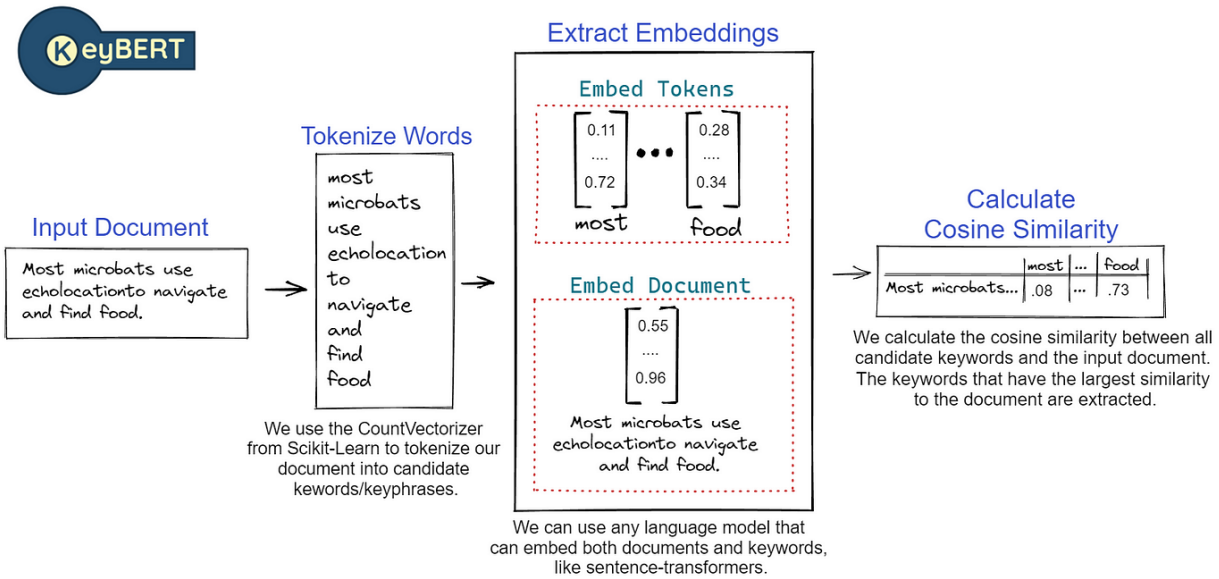

Fig. 1 – *Pipeline* de la méthode `keybert` (GROOTENDORST, 2020).

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

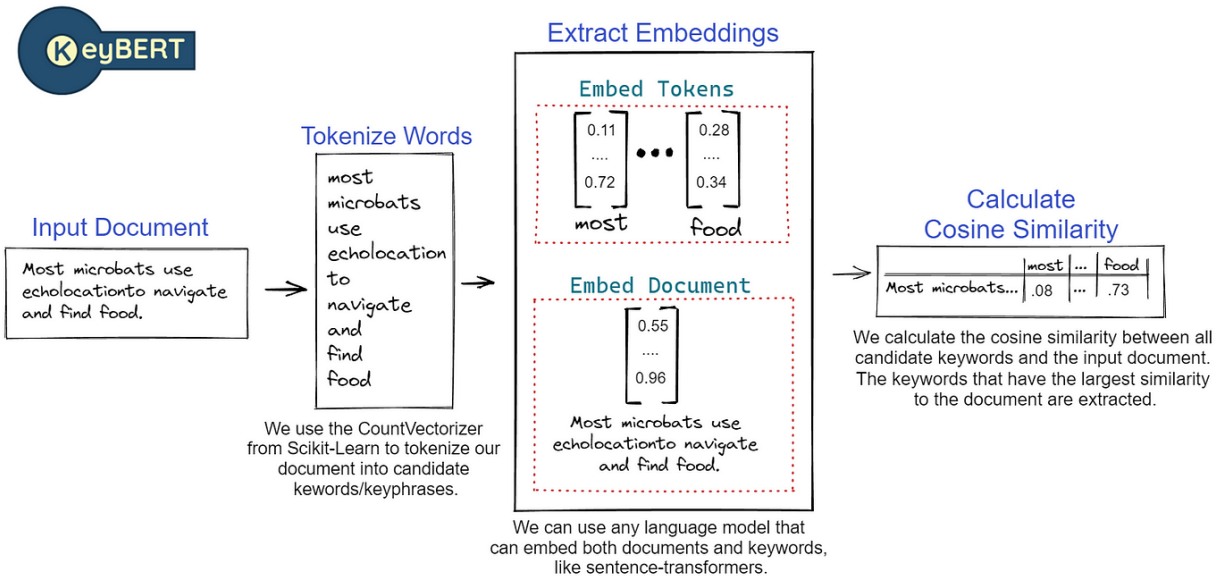

Fig. 1 – *Pipeline* de la méthode `keybert` (GROOTENDORST, 2020).

Maximal Marginal Relevance (MMR)

- paramètre de diversification des résultats
- basé sur la similarité cosinus

`use_mmr=True, diversity=[0-1]`

le degré de diversité entre 0 et 1

Maximal Marginal Relevance (MMR)

- paramètre de diversification des résultats
- basé sur la similarité cosinus

`use_mmr=True, diversity=[0-1]`

le degré de diversité entre 0 et 1

Liste des phrases-clés extraites avec `keybert`

Liste des mots vides appliquées : `spaCy`

Phrase-clé	Score
scientifique planches reproduction	0.5093
postérieure cordon postérieur	0.5078
cervico dorsale	0.465
sillon postérieur corne	0.4644
région cervicale figure	0.4572
cirrhose cancer primitif	0.4355
altération cellules ganglionnaires	0.4032
anatomie pathologique moëlle	0.3931
lcucocythcs substance granuleuse	0.3474
complètement détruite	0.334

(a) Corpus « Charcot ».

Phrase-clé	Score
magnétisme applicable horticulture	0.7012
droite corps envahie	0.567
action nitrite amyle	0.5114
trouve sabbat fans	0.4422
centimètres rotule circonférence	0.4194
mère attaques hystérie	0.4148
chloral décembre règles	0.4038
iconographie photographique salpetriere	0.3977
poitrine apparent 11	0.3388
hystérogènes description attaques	0.332

(b) Corpus « Autres ».

Table 1 – Liste de dix phrases-clés les plus pertinentes selon `keybert` dans les deux corpus.

`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

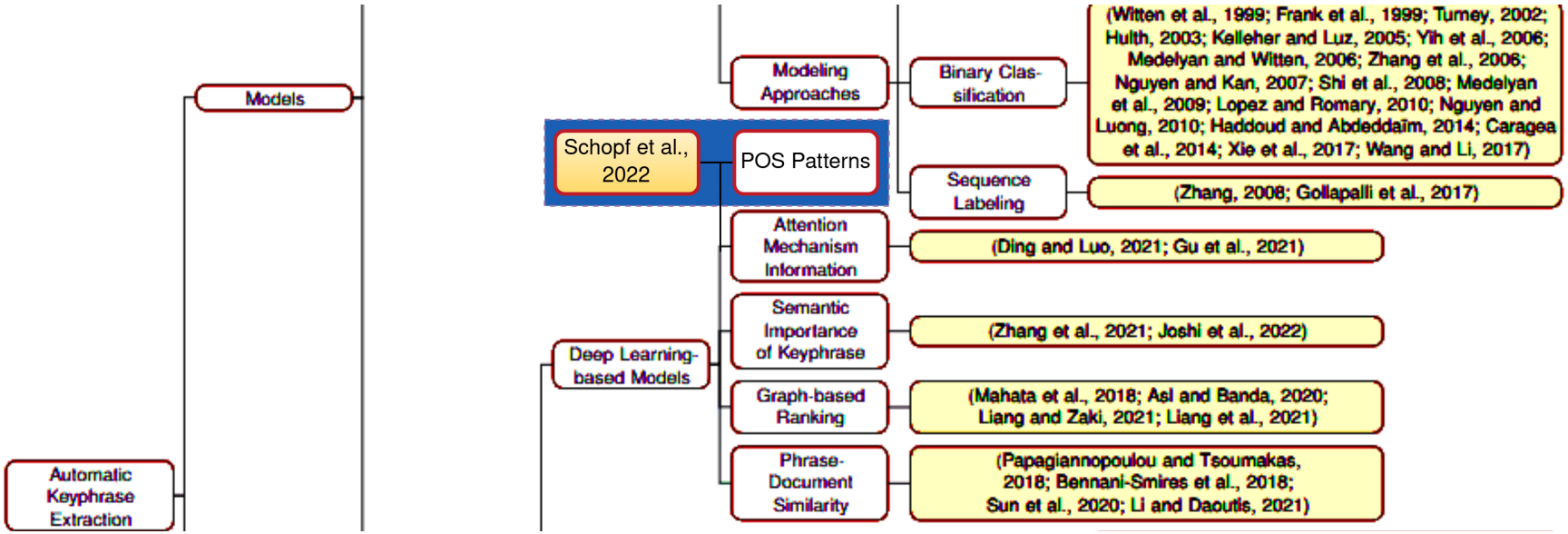

Fig. 2 – Extrait de l'état de l'art sur l'extraction des mots-clés, adapté de XIE et al. (2023)

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

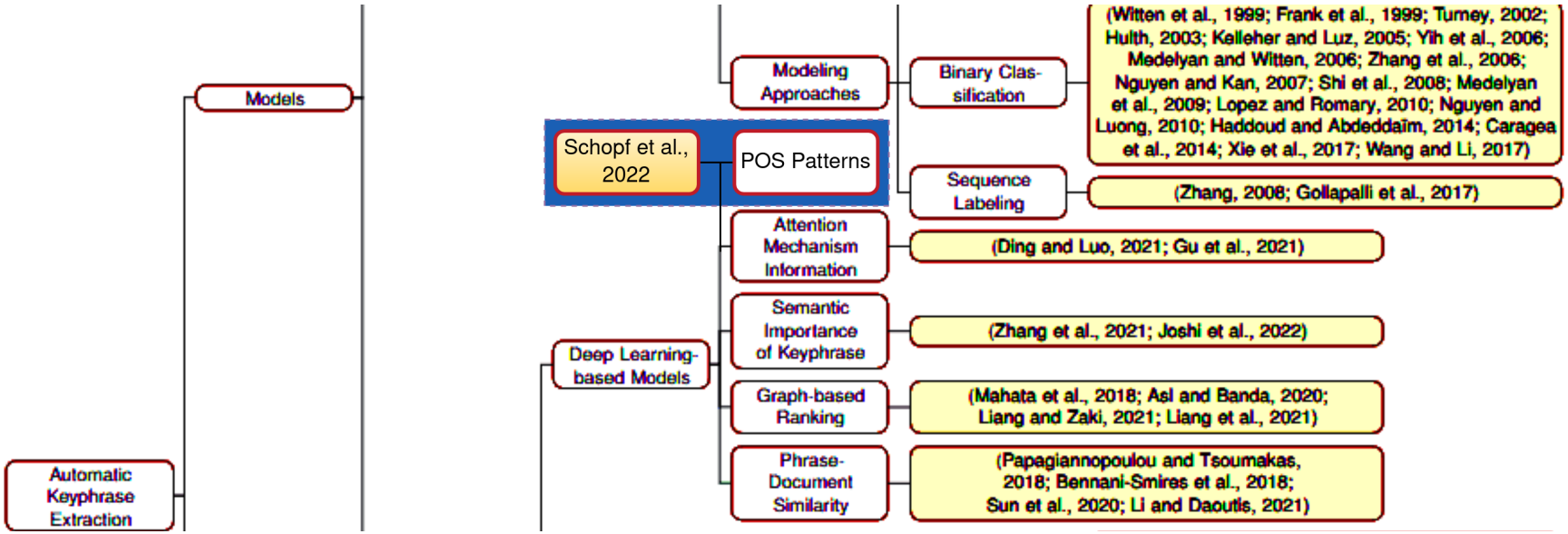

Fig. 2 – Extrait de l'état de l'art sur l'extraction des mots-clés, adapté de XIE et al. (2023)

`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

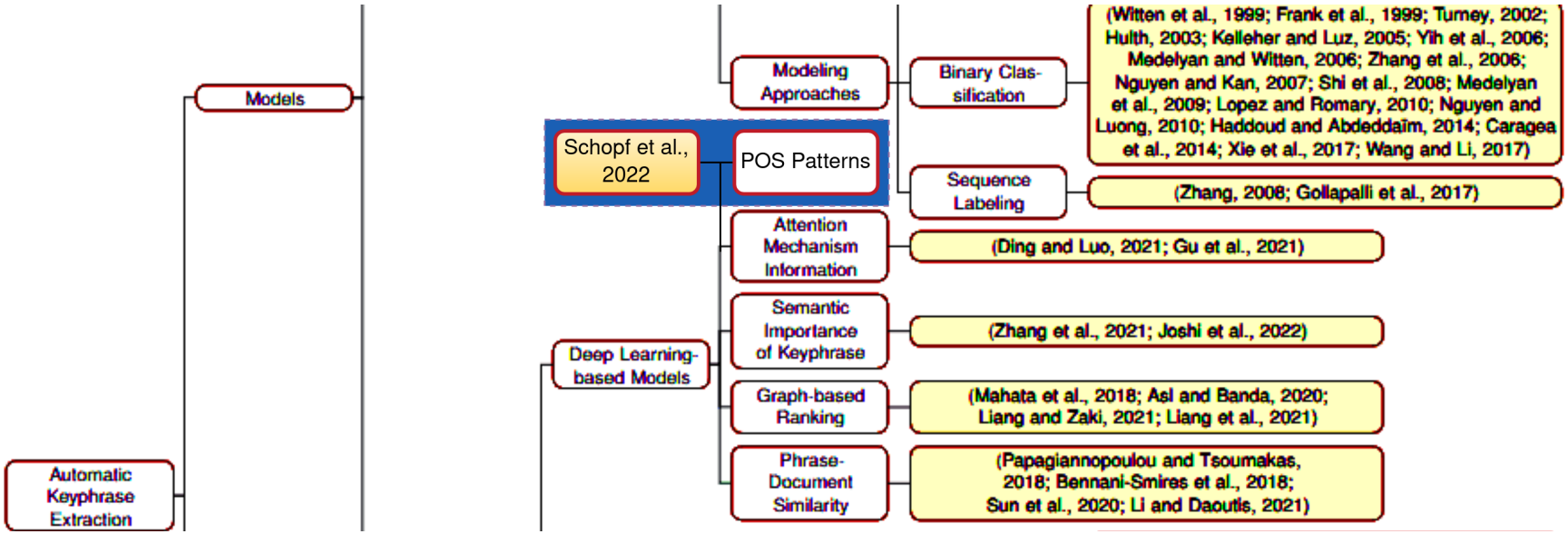

Fig. 2 – Extrait de l'état de l'art sur l'extraction des mots-clés, adapté de XIE et al. (2023)

`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

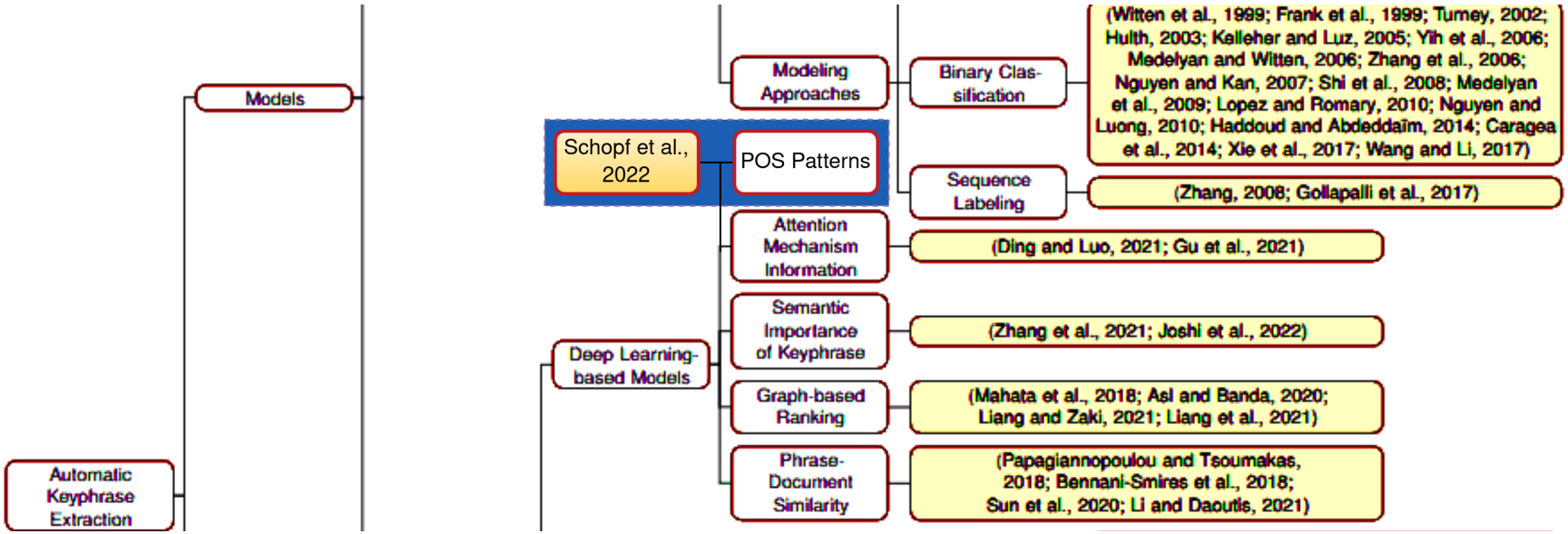

Fig. 2 – Extrait de l'état de l'art sur l'extraction des mots-clés, adapté de XIE et al. (2023)

Fonctionnement de la méthode *PatternRank*

1. entrée : un seul document texte tokenisé
2. étiquetage des tokens avec les balises POS
3. sélection des phrases-clés candidates correspondant au modèle POS
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

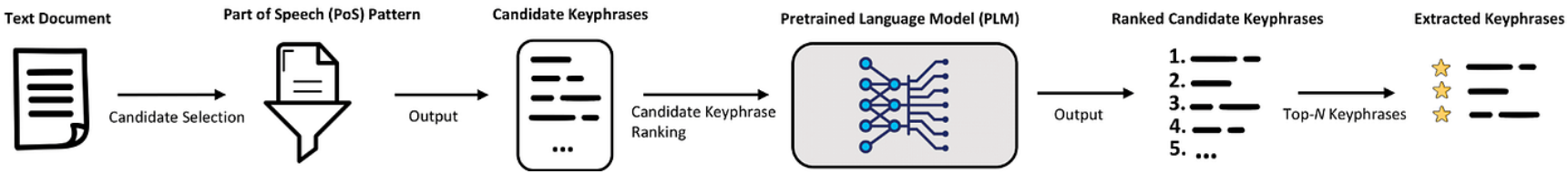

Fig. 3 – *Workflow* de la méthode *PatternRank* (SCHOPF et al., 2022).

Fonctionnement de la méthode *PatternRank*

1. entrée : un seul document texte tokenisé
2. étiquetage des tokens avec les balises POS
3. sélection des phrases-clés candidates correspondant au modèle POS
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

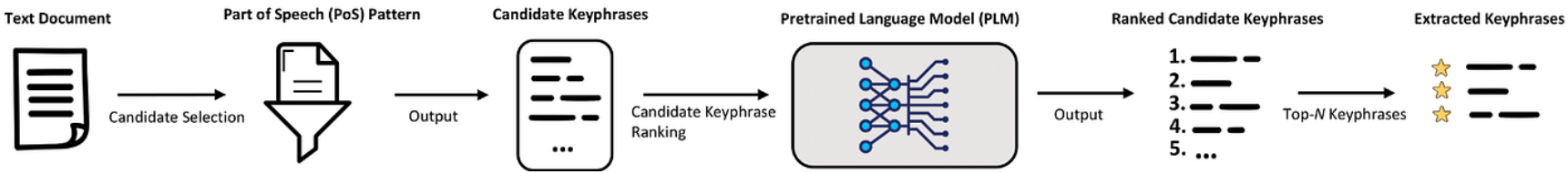

Fig. 3 – *Workflow* de la méthode *PatternRank* (SCHOPF et al., 2022).

Fonctionnement de la méthode *PatternRank*

1. entrée : un seul document texte tokenisé
2. étiquetage des tokens avec les balises POS
3. sélection des phrases-clés candidates correspondant au modèle POS
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

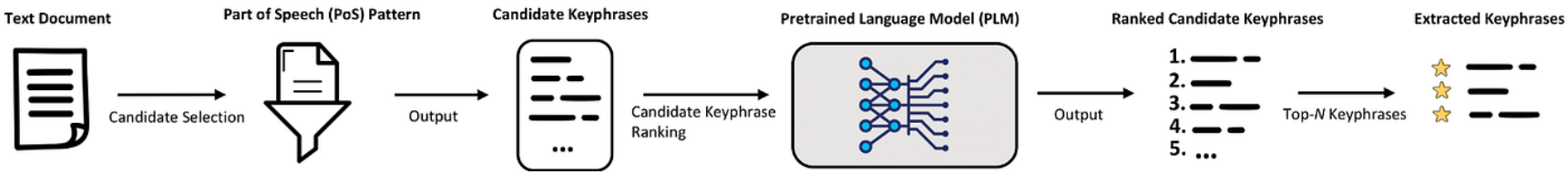

Fig. 3 – *Workflow* de la méthode *PatternRank* (SCHOPF et al., 2022).

Fonctionnement de la méthode *PatternRank*

1. entrée : un seul document texte tokenisé
2. étiquetage des tokens avec les balises POS
3. sélection des phrases-clés candidates correspondant au modèle POS
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

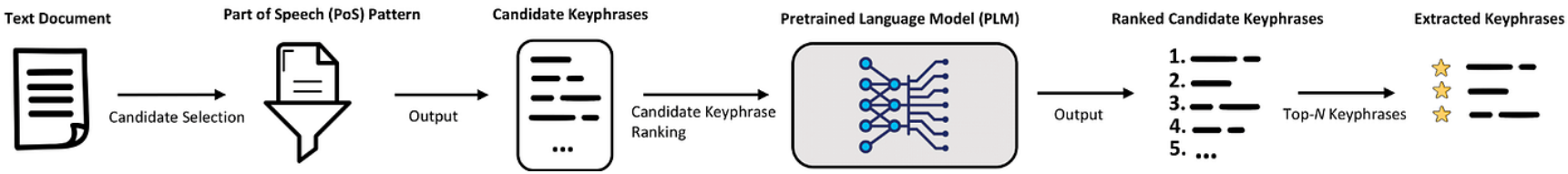

Fig. 3 – *Workflow* de la méthode *PatternRank* (SCHOPF et al., 2022).

Fonctionnement de la méthode *PatternRank*

1. entrée : un seul document texte tokenisé
2. étiquetage des tokens avec les balises POS
3. sélection des phrases-clés candidates correspondant au modèle POS
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

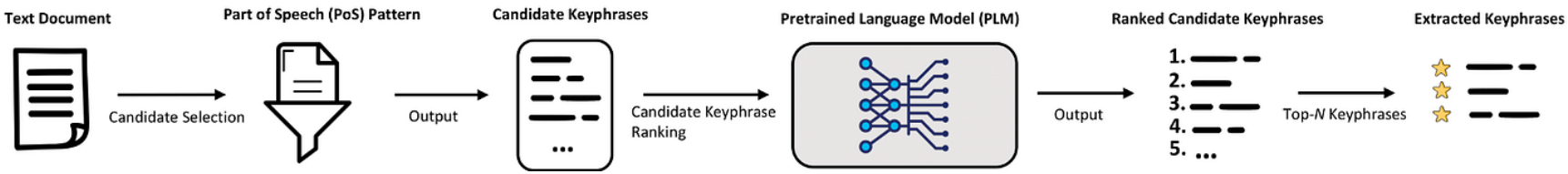

Fig. 3 – *Workflow* de la méthode *PatternRank* (SCHOPF et al., 2022).

Fonctionnement de la méthode *PatternRank*

1. entrée : un seul document texte tokenisé
2. étiquetage des tokens avec les balises POS
3. sélection des phrases-clés candidates correspondant au modèle POS
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

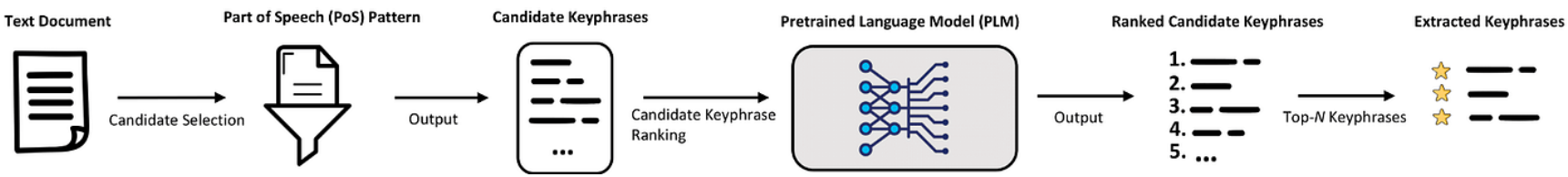

Fig. 3 – *Workflow* de la méthode *PatternRank* (SCHOPF et al., 2022).

Liste des phrases-clés avec `keyphrase-vectorizers`

Phrase-clé	Score
paroi épaissie	0.9276
histologie fine	0.9193
tissu gingival	0.9179
21c leçon	0.9162
travaux récents	0.9152
entrecroisement dos pyramides	0.9145
érysipèle périodique annuel	0.9135
cicatriciel	0.9118
fibromes	0.9109
affeclions	0.9091

(a) Corpus « Charcot ».

Phrase-clé	Score
trois lieues loing	0.9369
detfeins pernicieux	0.9292
21 mat	0.9289
attaques syncopales2	0.9278
accufoit	0.9255
diapason cataleptise	0.9252
membre faidt	0.9245
anes	0.9242
toutesfois	0.9235
demi culbute	0.9217

(b) Corpus « Autres ».

Table 2 – Liste des dix phrases-clés les plus pertinentes selon `keyphrase-vectorizers` dans les deux corpus.

Utilisation des librairies `keybert` et `keyphrase-vectorizers`

Ressources en ligne :

- Lien Google Colab

pré-requis :

- bonne connexion Internet
- mémoire RAM suffisante

- Dépôt GitHub

Utilisation des librairies `keybert` et `keyphrase-vectorizers`

Ressources en ligne :

- Lien Google Colab
  pré-requis :
  - bonne connexion Internet
  - mémoire RAM suffisante
- Dépôt GitHub

Utilisation des librairies `keybert` et `keyphrase-vectorizers`

Ressources en ligne :

- Lien Google Colab
  pré-requis :
  - bonne connexion Internet
  - mémoire RAM suffisante
- Dépôt GitHub

Utilisation des librairies `keybert` et `keyphrase-vectorizers`

Ressources en ligne :

- Lien Google Colab
  pré-requis :
  - bonne connexion Internet
  - mémoire RAM suffisante
- Dépôt GitHub

Utilisation des librairies `keybert` et `keyphrase-vectorizers`

Ressources en ligne :

- Lien Google Colab
  pré-requis :
  - bonne connexion Internet
  - mémoire RAM suffisante
- Dépôt GitHub

Passage à l'échelle

Pour traiter de grands corpus, il existe la possibilité de demander l'accès à la plateforme technologique MESU, hébergée par SACADO (Service d'Aide au Calcul et à l'Analyse de Données).

Elle est composée d'un supercalculateur, d'un environnement de virtualisation et d'un système de stockage de données.

Passage à l'échelle

Pour traiter de grands corpus, il existe la possibilité de demander l'accès à la plateforme technologique MESU, hébergée par SACADO (Service d'Aide au Calcul et à l'Analyse de Données).

Elle est composée d'un supercalculateur, d'un environnement de virtualisation et d'un système de stockage de données.

Passage à l'échelle

Pour traiter de grands corpus, il existe la possibilité de demander l'accès à la plateforme technologique MESU, hébergée par SACADO (Service d'Aide au Calcul et à l'Analyse de Données).

Elle est composée d'un supercalculateur, d'un environnement de virtualisation et d'un système de stockage de données.

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.

Broussolle, E., J. Poirier, F. Clarac et J.-G. Barbara (2012). Figures and institutions of the neurological sciences in Paris from 1800 to 1950. Part III : Neurology. In : *Revue Neurologique* 168.4, p. 301-320.

Camargo, C. H. F., L. Coutinho, Y. Correa Neto, E. Engelhardt, P. Maranhão Filho, O. Walusinski et H. A. G. Teive (2024). Jean-Martin Charcot : the polymath. In : *Arquivos de Neuro-psiquiatria* 81, p. 1098-1111.

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.

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).

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).

Grootendorst, M. (2020). *KeyBERT : Minimal keyword extraction with BERT*. Version v0.3.0 (*voir pp. 9-12*).

Références II

Koehler, P. J. (2013). Charcot, La Salpêtrière, and Hysteria as Represented in European Literature. In : *Progress in Brain Research* 206, p. 93-122.

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.

Marmion, J.-F. (2015). *Freud et la psychanalyse*. Sciences Humaines.

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.

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.

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. 16-25*).

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

Références III

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.

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. 16-19*).