

Extraction de termes significatifs :
de la représentation de documents
au résumé automatique.

Pascal Cuxac Léo Gaillard
CNRS - INIST

pascal.cuxac@inist.fr

leo.gaillard@inist.fr



L'extraction de termes

2 familles de méthodes :

- Avec une ressource contrôlée :
 - Projetée une terminologie sur un texte
 - Associe des termes d'une terminologie à des textes
- Sans ressource :
 - Extrait des termes «significatifs» d'un texte

But :

Recherche d'information
Représentation vectorielle de documents
...Résumé automatique ?...

Le résumé automatique

2 familles de méthodes :

- Le résumé par extraction :
Sélectionne des fragments de textes sans les modifier.
- Le résumé par abstraction :
Génère un texte en reformulant et condensant le texte initial

But :

Décrire un document (un corpus) de façon synthétique afin d'appréhender rapidement et facilement son contenu, de l'indexer en évitant les termes superflus, et réduire la taille de sa représentation vectorielle.

Extraction de termes

TEEFT : extraction de termes à partir du **texte intégral**
(Méthode PoS)

(basé sur NLTK et Topia + Filtrage en fonction de la distribution des poids des termes)

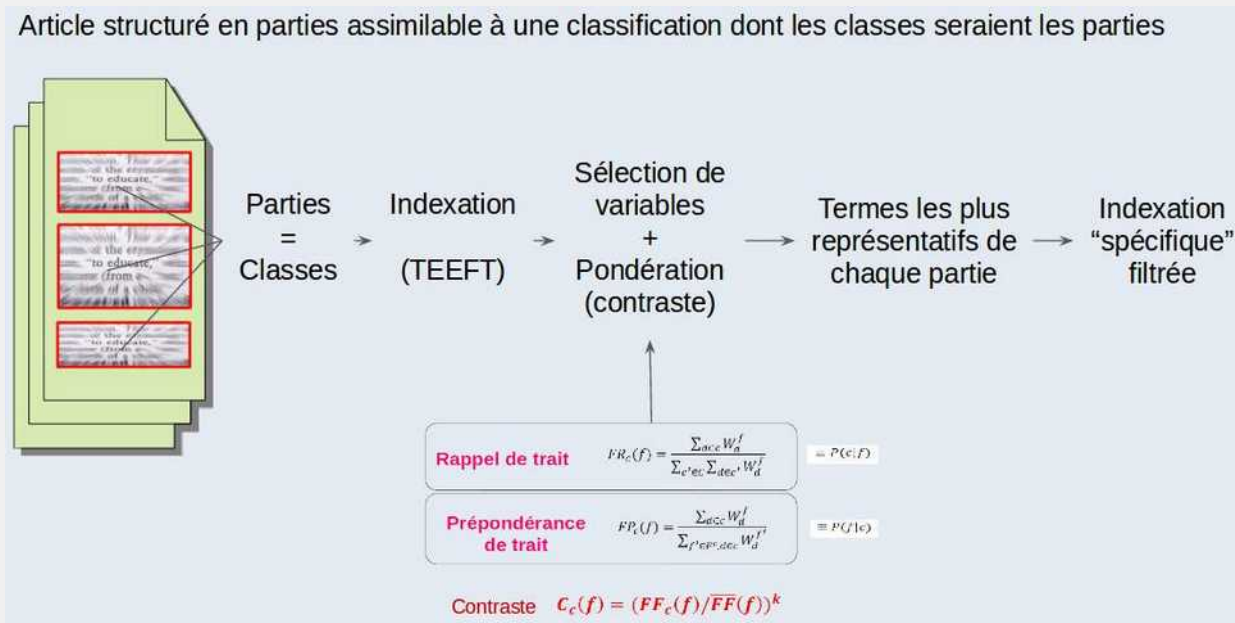


Teeft

nodule ;
bone nodule formation ;
osteoblast ;
hydroxy radical ;
anova ;
bone formation ;
non-coated well ;
extracellular matrix ;
rat calvarial cell ;
bone nodule area ;
bone cell ;
noncollagenous protein ;
blot analysis ;
goat igg fraction ;
nitrocellulose filter ;
cell-binding domain

Extraction de termes

Skeeft : Utiliser la structuration du document pour améliorer l'indexation par extraction de termes



Plia Hauk, MD, Katrin Friedl, Klaus Kaufmehl, MD, Radvan Urbanek, MD, and Johannes Forster, MD
From University Children's Hospitals, Freiburg, Germany, and Vienna, Austria

to investigate the risk of life-threatening reactions to future stings, we sequentially challenged 113 children (aged 2 to 17 years) allergic to insect stings with a sting by the relevant insect. The time interval between the challenges varied from 2 to 6 weeks. The history of the index stings was a large local reaction (LR) in 16% and a systemic reaction (SR) in 84% of the test subjects. On the first challenge, 76% had a normal LR, 11% a large LR, and 13% an SR. On the second challenge, 78% of the children had a normal LR, 5% a large LR, and 17% an SR. Thirty-nine of the untreated children were exposed to a field sting during the subsequent 3-year follow-up period. In comparison with other diagnostic evaluations such as skin-prick tests, determinations of specific IgE and IgG antibodies, and single-sting exposure, the dual sting challenge scheme appears to be the best predictor of reactions to subsequent stings. It also appears to be helpful in selecting patients with an uncertain sensitization status for venom immunotherapy. (J PEDIATR 1995;126:185-90)

[illegible]

Extraction de termes

Méthode	Rappel	Précision	F-mesure
TopicRank	0.11	0.18	0.14
Keyterm	0.25	0.21	0.23
SingleRank	0.06	0.09	0.07
Kea	0.14	0.21	0.17
KPMiner	0.17	0.22	0.19
Termostat	0.24	0.30	0.27
Teeft	0.23	0.20	0.21
Skeeft	0.21	0.32	0.25

TAB. 1 – *Comparaison des performances de Skeeft sur un corpus test indexé manuellement*

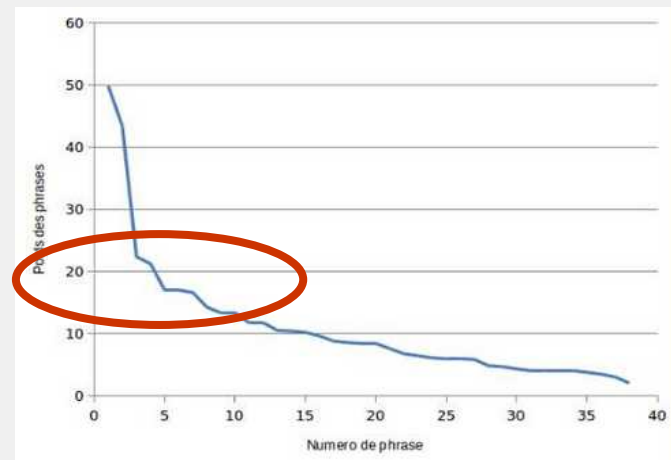
Résumé par extraction

Génération d'un résumé automatique par extraction de phrases

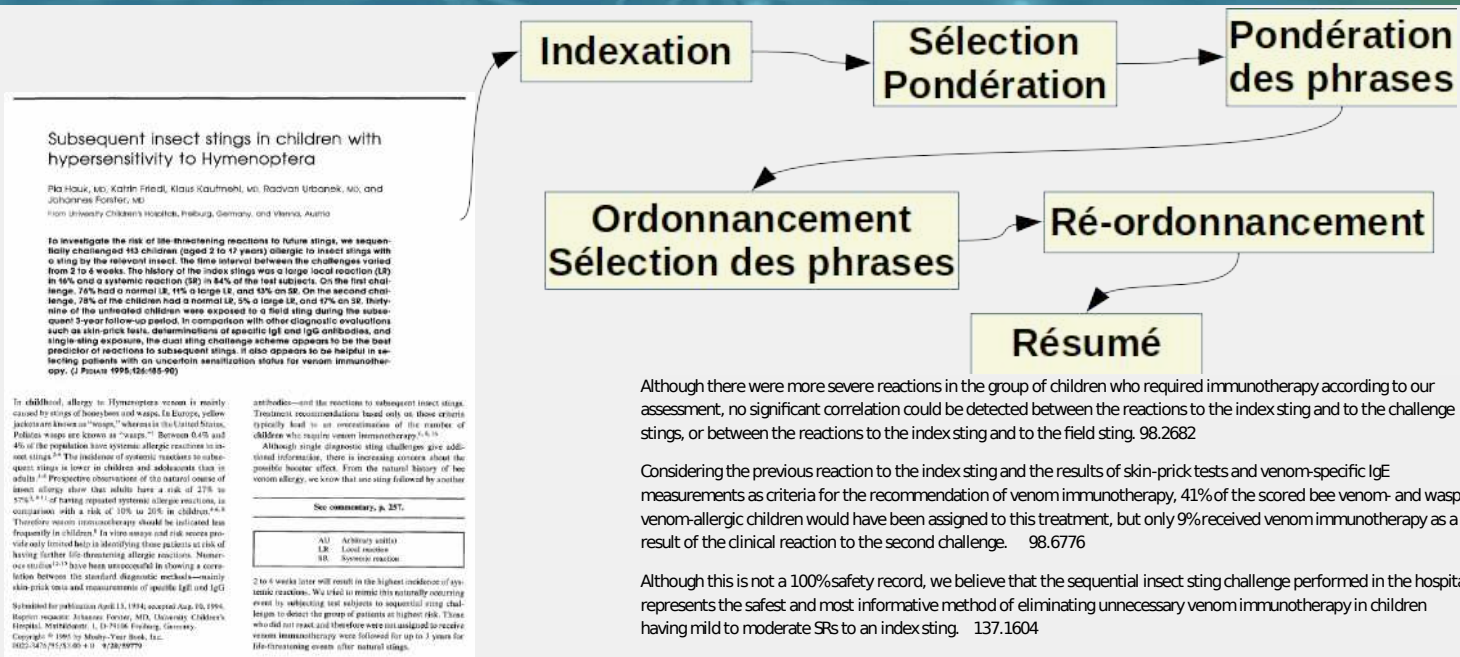
- Sélection / Pondération des termes
- Pondération des phrases
- Ordonnancement des phrases par pondération décroissante / Sélection automatique des « n » phrases les plus importantes
- Ré-ordonnancement par ordre d'apparition
- Résumé final

résumé automatique ?

Répartition du poids des phrases



Résumé par extraction



Résumé par extraction

Evaluation sur SciSumm

Rouge-2 : 0.169

(Best 0.329 abstract / 0.171 résumé humain)

Problèmes :

Données utilisateurs :
word, txt ou au mieux pdf

Obtention du xml :
problèmes liés à grobid
xml editeur (disponibilité ; qualité...)

Traitements de xml hétérogènes

Kokil Jaidka et al., 2019 The CL-SciSumm Shared Task 2018: Results and Key Insights, <https://arxiv.org/abs/1909.00764>

Sotaro Takeshita et al., 2024, ROUGE-K: Do Your Summaries Have Keywords? <https://arxiv.org/abs/2403.05186>

Al Saied et al. 2018. Automatic summarization of scientific publications using a feature selection approach. International Journal on Digital Libraries, 19(2/3), 203–215.
<https://doi-org.in2p3.bib.cnrs.fr/10.1007/s00799-017-0214-x>

Vous avez dit LLM ?

Génération d'un résumé automatique par abstraction

Problèmes possibles :

- hallucinations possibles
- taille de la fenêtre de contexte (peut imposer une fragmentation du texte [map-reduce])
- cohérence difficile à maintenir
- couts d'infrastructure pouvant être élevé (GPU)
- temps d'exécution

Vous avez dit LLM ?

Génération d'un résumé automatique par abstraction

Un service basé sur un «petit» LLM

- **Bart-large-CNN**, obtenu après fine-tuning du modèle Bart-large sur le jeu de données CNN-dailymail
- encodeur/décodeur (Transformer)
- Test sur SciSumm...le meilleur résultat (pour un «petit» llm)

Lewis, M. et al. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. CoRR, abs/1910.13461.
<http://arxiv.org/abs/1910.13461>

Yasunaga, M. et al. (2019). ScisummNet: A Large Annotated Corpus and Content-Impact Models for Scientific Paper Summarization with Citation Networks. arXiv preprint arXiv:1909.01716.
<https://arxiv.org/abs/1909.01716v3>

SciSumm : <https://huggingface.co/datasets/usamahhanif719/scisumm>

Vous avez
dit LLM ?

Génération d'un résumé automatique par abstraction

Web service en production :

<https://services.istex.fr/resume-automatique-dun-article-scientifique/>

ROUGE-2 score : 0.28

Bien meilleur que skeeft mais dépendant de la langue...et un peu plus lourd
(mais tourne sans GPU!)

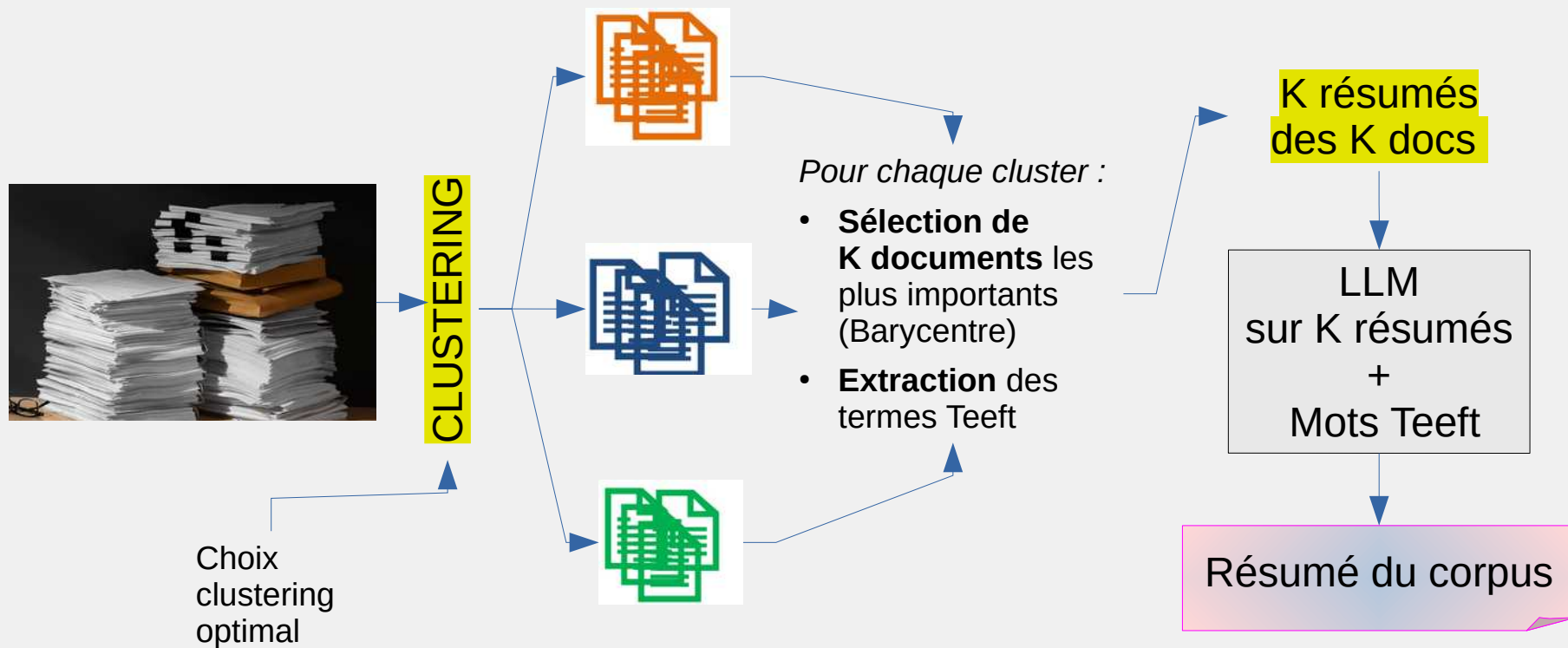
Modèle multilingue peu performant et plus gourmand

Et si on résumait un corpus de documents ?

Résumé d'un ensemble de documents pour :

- Compléter les métadonnées associées à un corpus spécialisé par exemple
- Avoir une idée de l'homogénéité d'un gros corpus
- Décrire de façon simple le contenu de clusters d'une classification non supervisée
- ...

Et si on résumait un corpus de documents ?



Et si on résumait un corpus de documents ?

Comment évaluer le résultat ? De nouvelles méthodes à inventer....

- Validation humaine
- Validation automatique :
 - Calcul de la distance du résumé généré à l'isobarycentre du corpus

Conclusion

Résumé automatique d'un document bien traité actuellement

Avantages de Skeeft :

- Méthode légère
- Indépendant de la langue du document

Défauts de Skeeft :

- Besoin de connaître la structure du document
- Supplanté par les LLM

Résumé automatique de corpus encore à investiguer

- Difficile à traiter tel quel par un LLM
- Méthode d'évaluation à étudier
- Corpus d'évaluation à construire

Extraction de termes significatifs :
de la représentation de documents
au résumé automatique.

Merci de votre attention



ISTEX

Pascal Cuxac

Léo Gaillard

CNRS - INIST

pascal.cuxac@inist.fr

leo.gaillard@inist.fr

Séminaire ALIMining
Toulouse 29-30 septembre 2025