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



2 familles de méthodes :

- Avec une ressource contrôlée :
 Projette 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.

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)

 H_2O_2 -derived free radicals treated fibronectin substratum reduces the bone nodule formation of rat calvarial osteoblast

Hiroshi Suzuki ^a, Mitsuo Hayakawa ^b, Kihei Kobayashi ^a, Hisashi Takiguchi ^b, Yoshimitsu Abiko ^{b,*}

Department of Complete Denture Prosthodontics, Nihon University School of Dentistry at Matsudo, 2-870-1, Sakaecho-Nishi, Matsudo, Chiba 271, Japan

^b Department of Biochemistry, Nihon University School of Dentistry at Matsudo, 2-870-1, Sakaecho-Nishi, Matsudo, Chiba 271, Japan

Received 28 October 1996; accepted 21 April 1997

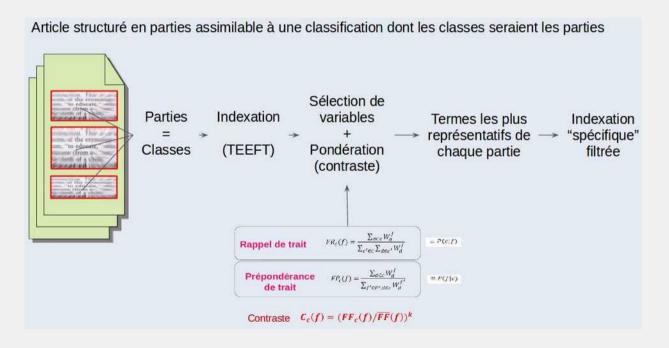
Abstract

Fibronectin (FN) is involved in various cellular activities such as adhesion, proliferation and migration as a substratum. Since the metabolic turnover of FN is much slower than other cellular components, it may be affected by the oxygen free radicals produced in the aging process. However, the effect of oxygen free radicals on FN as substratum in bone formation has not been well characterized. The objective of this study was to examine the effect on the bone forming activity of osteoblasts using an oxygen free radical treated FN substratum in vitro (H₂O₂-Cu²⁺ system). SDS-PAGE, Western blotting and immuno-blotting analysis revealed that FN was degradated and/or modified by H₂O₂-Cu²⁺ (·OH) treatment. Bone nodule formation per well was examined for total number, total area and area per nodule, which data were then compared between non-coated and FN-coated, and between FN-coated and ·OH treated FN-coated between the romation in the FN-coated was significantly greater than in the non-coated. Furthermore, bone nodule formation in ·OH treated FN-coated was significantly less than that of FN-coated. These findings suggested that FN plays important roles in osteoblast activity and that FN substratum

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

https://api.istex.fr/ark:/67375/6H6-Q3G9XGNS-T/fulltext.pdf

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



Subsequent insect stings in children with hypersensitivity to Hymenoptera

Pia Hauk, MD, Katrin Friedl, Klaus Kaufmehl, MD. Radvan Urbanek, MD, and Johannes Forster, MD

From University Children's Hospitais, 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 (IR) in 16% and a systemic reaction (SR) in 84% of the test subjects. On the first challenge, 76% had a normal IR, 11% a large IR, and 13% an SR. On the second challenge, 78% of the children had a normal IR, 5% a large IR, and 17% an SR. Initynine 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 igle and igle 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 Pricar 1995; 126:165-90)

The control of the co

Text rous beyond rous text-cockl. res(ELSH-/os/text-cyc-beyond--/os-beyond-sed-risos-text-ris

reactives to the language of the contractive contracti

was person called and, allergy to represent the control of control to those of management and users. In future, setting partitions the control to the contro

extension of the depreciation of the second of the second

respectably prist being we perform with perform with perform with principle has not any seens that principle accounts assumed to the foreign act constitutions of 0.1, 1, 10, and 300 counts got to entire account of the foreign action of the foreign accounts action of the foreign action

re notion tillaction trained in Apprentition is notice (non-tillaction) and the following interesting tillaction to be a supposed using the following interest positive delayable to be a supposed using non-tillaction and the following integrated large (E. on 18 concepts).

The control of the

exponent grantform on approach by the edition control consists of the opinists, and informed counts and integral from the parents, in particular other report to the parents debugs of the control of the consists of the cons

constitutivities and productive and the second of the seco

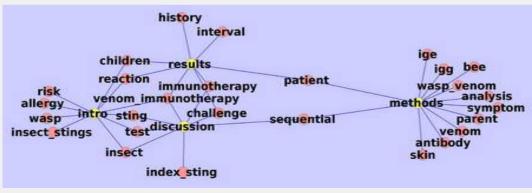
comprehensive characteristics of patients are time to recommend religiously. The This conservation had been different types of jobs play must be be significant difference in the section of the section

ons mention-testing can improve account the many control of the co

compared to the control of the contr

Constitution of the Compression of the Compression





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

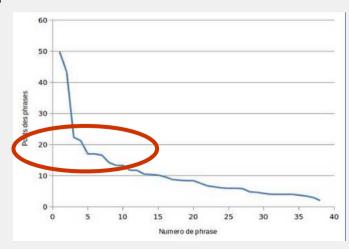
Résumé par extraction

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 finales um é

automatique?





Résumé par extraction

Subsequent insect stings in children with hypersensitivity to Hymenoptera

Pla Hauk, No, Katrin Friedl, Klaus Kaufmehl, vo. Radvatt Urbanek, No, and Johannes Forster, No.

From University Children's Hospitals, Preiburg, Germany, and Vienna, Austria

In investigate the risk of the Private International Control States (1997) and the Private International Control Control States (1997) and the Private International Control States (1997) and the International Control Contr

In childhood, allerey to Hymeroptera verson is mainly carried by stings of hones have and wasne. In Farons, wellow jackota are known on "woogs," whereas in the United States, Polistes wasps are known as "warps." Between 0.4% and 4% of the population have systemic affergic reactions to insort stings 24 The incidence of systemic machines to subsequest stings is lower in children and adolescents than in adults.1-4 Prospective observations of the natural course of insect allergy show that adults have a risk of 27% to 57% h 11 of having repeated systems: allergic reactions, in comparison with a risk of 10% to 20% in children.*4.8 Therefore watern immuniciberage should be indicated less frequently in children. In vitro assays and risk scores provide only limited belo in identifying those patients at risk of having further life threatening affernic reactions. Numeroce studies 12-13 have been unapocessful in showing a correlation between the standard digentatic methods-mainly skin-prick tests and measurements of specific left and leG

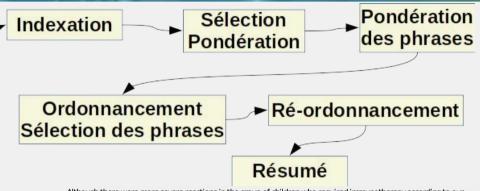
Submitted for problems and Ed. 1, 1994, occupied Aug. 10, 1994. Ropers reconstructures Forster, MD, University Children's Reconstructures and Parlies Forstering, Geometry, Copyright # 1995 by Monty-Year Book. Sta. 1985; 2020. 455; 400 4.11 42-200-407-70. antibodies—and the reactions to subsequent intext stings. Treatment recommendations based only on those criteria typically lead to an overestimation of the number of children who require venum immentatorages (s. 18).

Although single diagnosts sting shallenges give additional information, there is increasing covern about the possible become effect. From the natural biscary of bewomen effergy, we know that are sting followed by another

See commentary, p. 257.

AU Achitrary unit(s)
LR Local reaction
IR Systems reaction

2 to 4 weeks later will could in the highest incidence of systemic reactions. We below to remis this naturally occurring event by subjecting test subjects to sequential using challesigns to describe group of gattents or highest risk. The world of who did not react and therefore were an analyzed to receive vasous immunolistrary were followed for up to 3 years for life-formationing overast offer natural stings.



Although there were more severe reactions in the group of children who required immunotherapy according to our assessment, no significant correlation could be detected between the reactions to the index sting and to the challenge stings, or between the reactions to the index sting and to the field sting, 98.2682

Considering the previous reaction to the index sting and the results of skin-prick tests and venom-specific IgE measurements as criteria for the recommendation of venom immunotherapy, 41% of the scored bee venom- and wasp venom-allergic children would have been assigned to this treatment, but only 9% received venom immunotherapy as a result of the clinical reaction to the second challenge. 98.6776

Although this is not a 100% safety record, we believe that the sequential insect sting challenge performed in the hospital represents the safest and most informative method of eliminating unnecessary venom immunotherapy in children having mild to moderate SRs to an index sting. 137.1604

On the basis of the data presented, we suggest the following diagnostic and therapeutic procedures for children up to 16 years of age: Sensitized patients, identified by a positive skin-prick test result or specific IgE finding, who had only a large LR to the index sting, need neither a challenge sting nor venom immunotherapy. 104.307

Résumé par extraction

Evaluation sur SciSumm

Rouge-2: 0.169

(Best 0.329 abstract / 0.171 resumé humain)

automatique

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

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

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/usamahanif719/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

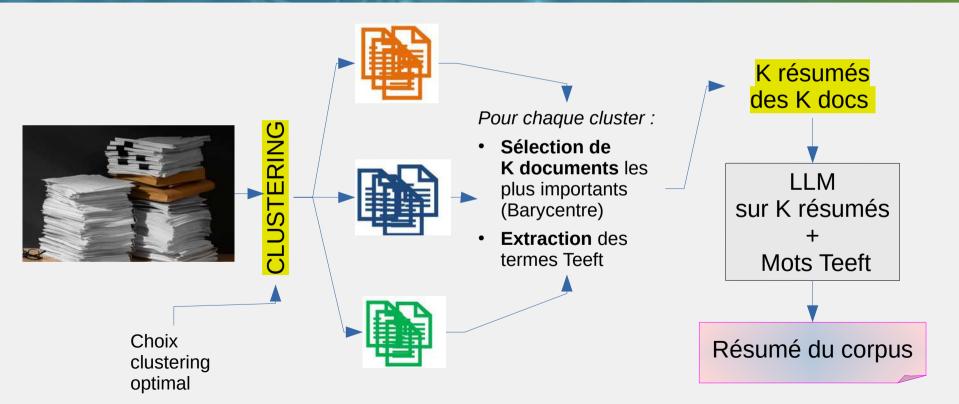
Et si on résumait un corpus de documents ?

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



Pascal Cuxac Léo Gaillard CNRS - INIST

pascal.cuxac@inist.fr

leo.gaillard@inist.fr



Séminaire ALIMining Toulouse 29-30 septembre 2025