



Chantier d'usage ISTEEX

Inria-Alpage

Équipe



Aazhar



**Patrice Lopez
(kermitt2)**



Laurent Romary

Les enjeux de la fouille de textes scientifiques

- e-science : exploitation du calcul intensif pour tirer profit de volumes massifs de données scientifiques
 - ➔ biologie, bioinformatique, génomique, physique, astronomie, sciences humaines
- Pour (Hey, 2009) : le début d'une quatrième révolution scientifique où l'usage des ordinateurs permettra la création de nouveaux concepts, idées, modèle et simulations, et une "renaissance" scientifique
- Fouille de textes scientifiques : rendre exploitable l'information contenu dans la littérature scientifique et technique
 - ➔ Amélioration des outils de recherche d'information,
 - ➔ Construction de bases de connaissances,
 - ➔ À plus long terme, production automatique d'hypothèses scientifiques (Evans, 2010)

Les blocages

ISTEX

- Droit d'accès à la littérature scientifique et technique
- Légalité de la fouille de texte (US/UK/JP vs FR/DE)
- Besoin de couverture, données à jour (+ 1,5M articles/an)

GROBID

- Difficulté d'exploitation du format PDF / pauvreté et incohérence des metadonnées
- *“the information that I can extract from an article, at least for me, is not quite the information I want”*, Shreejoy Tripathy (neuroscientifique)

Les défis scientifiques et techniques

- Gestion de la volumétrie : million de documents, milliards d'annotations, graphe avec milliard de noeuds
- Gestion de l'incertitude et du bruit: les meilleures techniques de fouille de textes font des erreurs, beaucoup d'erreurs...
- *Machine learning* : domaines et scénarios plus complexes et ouverts que le web marchand et les “add clicks”
- Prise en compte des méthodologies, opinions, réseaux, etc. propres à chaque discipline

Tirer profit de la fouille de textes dans ISTEK

- Un très grand volume de données afin de neutraliser le bruits, les erreurs d'extractions, etc.
- Décloisonner les disciplines: croiser des champs/ disciplines
- Travail de normalisation, nettoyage des données extraites
- Capturer le langage spécialisé

Objectifs du chantier d'usage ISTEEX

- Tester et passer à l'échelle nos propres annotateurs
- Combiner analyse sémantique multi-domaine et extraction spécialisée de nomenclature scientifique
- Monter en charge notre infrastructure d'analyse documentaire scientifique
- Illustrer l'intérêt d'ISTEEX pour la fouille de texte

Travail du chantier d'usage ISTEEX

1. **Annotations spécialisées en information scientifique et technique sur un grand échantillon** - 10 à 20% de l'ensemble du corpus ISTEEX
 - (N)ERD : entités connus / catégorisation
 - BIO-GROBID/BEAST : biotech
 - CHEMICAL-GROBID : chimie
 - GROBID-quantities : mesures physiques
2. Encodage TEI stand-off des annotations
3. Démonstrateur exploitant ces annotations basé sur notre infrastructure anHALytics

Approche

- Les approches par apprentissages automatiques dominant toutes les compétitions en extraction d'information de textes scientifiques

How to improve the performance of a ML application?

- Get better algorithm? +
- Get better features? +++
- Get better data? ++++++

(Leon Bottou)

Scores de correction actuelle en tâche d'extraction d'information

Information type	Accuracy
Entities	90-98%
Attributes	80%
Relations	60-70%
Events	50-60%

(A. McCallum)



Chiffres très optimistes pour des domaines et tâches bien établis (ex. biotechnologie) et des textes non bruités

Connaissances scientifiques et littérature académique

- Le langage spécialisé (terminologie, nomenclature, etc.) est le véhicule principal par lequel la connaissance technique et scientifique est représentée et transmise (Ahmad, 1996)
- Concepts connus vs entités nouvelles vs nomenclatures
- Un terme est l'expression d'un concept scientifique:
 - ➔ Principe Würsterien en terminologie : un concept est exprimé par un terme (pas de synonymie) et un terme ne peut référer qu'à un concept (pas de polysémie).
 - ➔ Pour le TALN et les applications ciblées : l'ambiguïté terminologique est constante, la confrontation des concepts, des domaines et des données nourrit le débat scientifique - c'est là que l'on trouve les nouvelles idées et que l'on peut faire émerger de nouvelles hypothèses intéressantes pour les chercheurs

Capturer les concepts connus : (N)ERD


- Identification et résolution des entités avec des bases de connaissances : Wikipedia, FreeBase (Wikidata)
 - Couverture immédiate : > 4M de “concepts”, ~150M formes lexicales (pour l’anglais)
 - Pas d’autres contraintes d’expertise humaine en ingénierie des connaissances
 - Pas de contrainte de domaine
 - Multilingue
 - Catégorisation suivant les catégories Wikipedia/type Freebase
- ➔ mais ne couvre que les entités connues!

(N)ERD

Commonness n'est pas suffisant : (Milne & Witten, 2008)

Depth-first search

From Wikipedia, the free encyclopedia



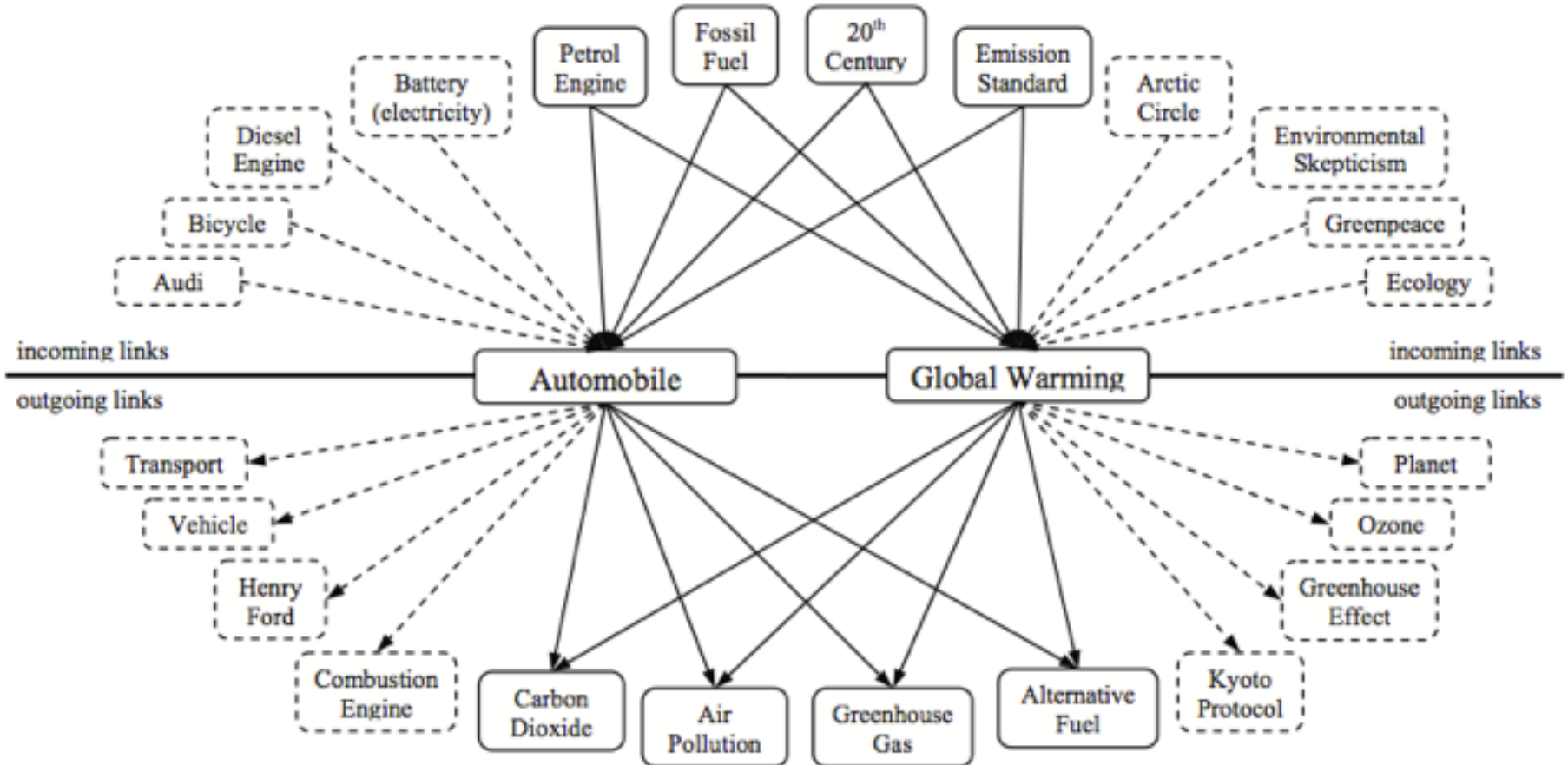
Depth-first search (DFS) is an **algorithm** for traversing or searching a **tree** **tree structure** or **graph**. One starts at the root (selecting some node as the root in the graph case) and explores as far as possible along each branch before **backtracking**.

Formally, DFS is an **uninformed search** that progresses by expanding the first child node of the search **tree** that appears and thus going deeper and deeper until a goal node is found, or until it hits a node that has no children. Then the search **backtracks**, returning to the most recent node it hadn't finished exploring. In a non-recursive implementation, all freshly expanded nodes are added to a **LIFO stack** for exploration.

sense	commonness
Tree	92.82%
Tree (graph theory)	2.94%
Tree (data structure)	2.57%
Tree (set theory)	0.15%
Phylogenetic tree	0.07%
Christmas tree	0.07%
Binary tree	0.04%
Family tree	0.04%
...	

NERD

Relatedness: (Milne & Witten, 2009)



Reconnaissance d'entités scientifiques liées à des nomenclatures

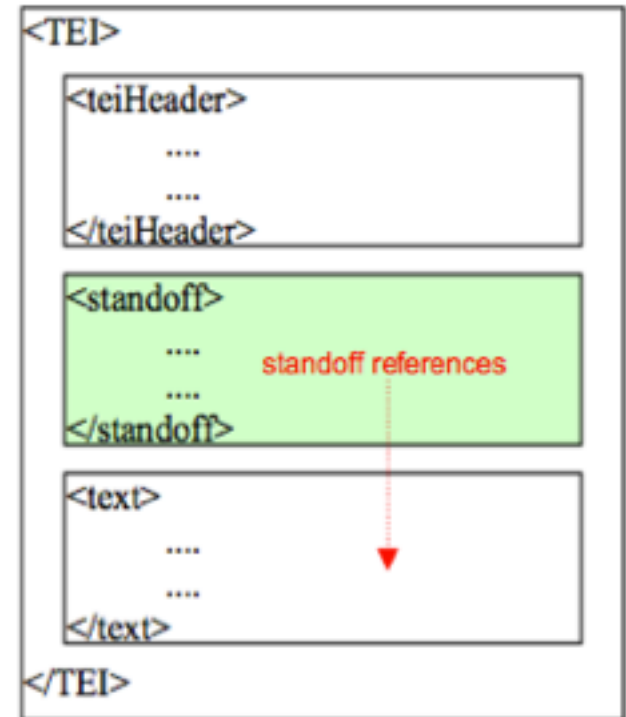
- Nomenclature: un système de nommage permettant de produire des termes et des concepts dans un champ particulier des sciences
- Terminologie et nomenclature = réduire l'impact de la langue naturelle
- Les entités ne peuvent toutes être énumérées par avance car les nomenclatures sont génératives.
 - ➔ PubChem du NIH par exemple reporte plus de 100 millions de formules chimiques dans le corpus brevets mondial
 - ➔ Central en chimie, biologie, astronomie, etc.
 - ➔ Pas de recherche professionnelle d'infos dans ces domaines sans prendre en compte ces termes

Travail du chantier d'usage ISTEEX

1. Annotations spécialisées en information scientifique et technique sur un grand échantillon - correspondant à 10 à 20% de l'ensemble du corpus ISTEEX
 - (N)ERD : entités connus / catégorisation
 - BIO-GROBID/BEAST : biotech
 - CHEMICAL-GROBID : chimie
 - grobid-quantities : mesures physiques
2. **Encodage TEI stand-off des annotations**
3. Démonstrateur exploitant ces annotations basé sur notre infrastructure anHALytics

Codage TEI d'annotations automatiques

- Persistance et réutilisation des annotations
- Nécessité de schémas détaillés d'annotations, agnostique -> TEI
- Annotations *standoff* en TEI avec offset
- Gestion de multiples annotations concurrentes, cumulables
- Positions préservées



Travail du chantier d'usage ISTEK

1. Annotations spécialisées en information scientifique et technique sur un grand échantillon - correspondant à 10 à 20% de l'ensemble du corpus ISTEK
 - (N)ERD : entités connus / catégorisation
 - BIO-GROBID/BEAST : biotech
 - CHEMICAL-GROBID : chimie
 - grobid-quantities : mesures physiques
2. Encodage TEI stand-off des annotations
3. **Démonstrateur exploitant ces annotations basé sur notre infrastructure anHALytics**

anHALytics

- Comment améliorer l'engagement des chercheurs français en faveur de l'Accès Libre et HAL?
- Parmi un ensemble d'actions complémentaires, l'idée de l'ADT AnHALytics est d'offrir, en complément du service d'archivage existant, **une plate forme analytique** permettant
 - ➔ la promotion de HAL comme observatoire de l'activité scientifique à différents niveaux de granularité
 - ➔ une motivation supplémentaire pour une politique de mandat de dépôt ambitieuse telle que celle de l'Inria

SOURCES

arXiv.org

PMC

HAL
archives-ouvertes.fr

OAI-PMH

PDF

TFI

XMI

ISTEX

...

Crossref

\$

DEDUPLICATION

KNOWLEDGE BASE

Entités de la recherche

GRAPHDB

TEI

INDEX

PERSISTANCE

PDF

OCR

XML/TEI

ASSETS

ANNEXES

ANNOTATEURS

OCR

GROBID

(N)ERD

KEYTERM

GROBID-QU

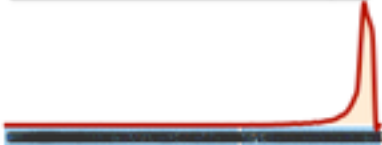
BIOGROBID

...

Demo: <http://traces1.saclay.inria.fr/anHALytics>

+ add new facet

publication_date



DD MM YYYY to DD MM YYYY ✓

subject-headers



keywords

- + physical sciences (872)
- + france (556)
- 160-160000 (100)

search term

Disamb./Expand

76,133 results - in 257 ms (server time)

ird-00968855 - Intercomparison of four remote-sensing-based ENERGY BALANCE methods to retrieve SURFACE EVAPOTRANSPIRATION and WATER STRESS of IRRIGATED FIELDS in SEMI-ARID CLIMATE
J. Chirouze et al. - 31.12.2014



ird-00968855

JOURNAL ARTICLES

Jonas Chirouze

G. Boulet

L. Jarlan

R. Fieuzal

J. C. Rodriguez

J. Ezzahar

S. Er Raki

G. Bigeard

O. Merlin

J. Garatuzza-Payan

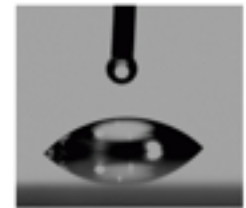
C. Watts

G. Chehbouni

Abstract: Instantaneous EVAPOTRANSPIRATION RATES and surface WATER STRESS levels can be deduced from REMOTELY SENSED SURFACE TEMPERATURE DATA through the SURFACE ENERGY budget. Two FAMILIES of methods can be defined: the contextual methods, where STRESS levels are scaled on a given IMAGE between hot/dry and cool/wet PIXELS for a particular VEGETATION cover, and single-pixel methods, which evaluate LATENT HEAT as the residual of the surface ENERGY BALANCE for one PIXEL independently from the others. Four models, two contextual (S-SEBI and a modified TRIANGLE method, named VIT) and two single-pixel (TSEB, SEBS) are applied over one GROWING SEASON (December-May) for a 4 KM x4 km IRRIGATED AGRICULTURAL AREA in the SEMI-ARID northern Mexico. Their performance, both at local and SPATIAL standpoints, are compared relatively to ENERGY BALANCE DATA acquired at seven locations within the area, as well as an UNCALIBRATED soil- vegetation-atmosphere transfer (SVAT) MODEL forced with local in situ DATA including observed IRRIGATION and RAINFALL amounts. STRESS levels are not always well retrieved by most models, but S-SEBI as well as TSEB, although slightly biased, show good performance. The drop in MODEL PERFORMANCE is observed for all MODELS when VEGETATION is

SURFACE ENERGY

Domains:
Engineering
conf: 0.88



“Surface energy” quantifies the disruption of intermolecular bonds that occurs when a surface is created. In the physics of solids, surfaces must be intrinsically less energetically favorable than the bulk of a material, otherwise there would be a driving force for surfaces to be created, removing the bulk of the material (see sublimation (chemistry)/sublimation). The surface energy may therefore be defined as the

Liens



Apache 2.0

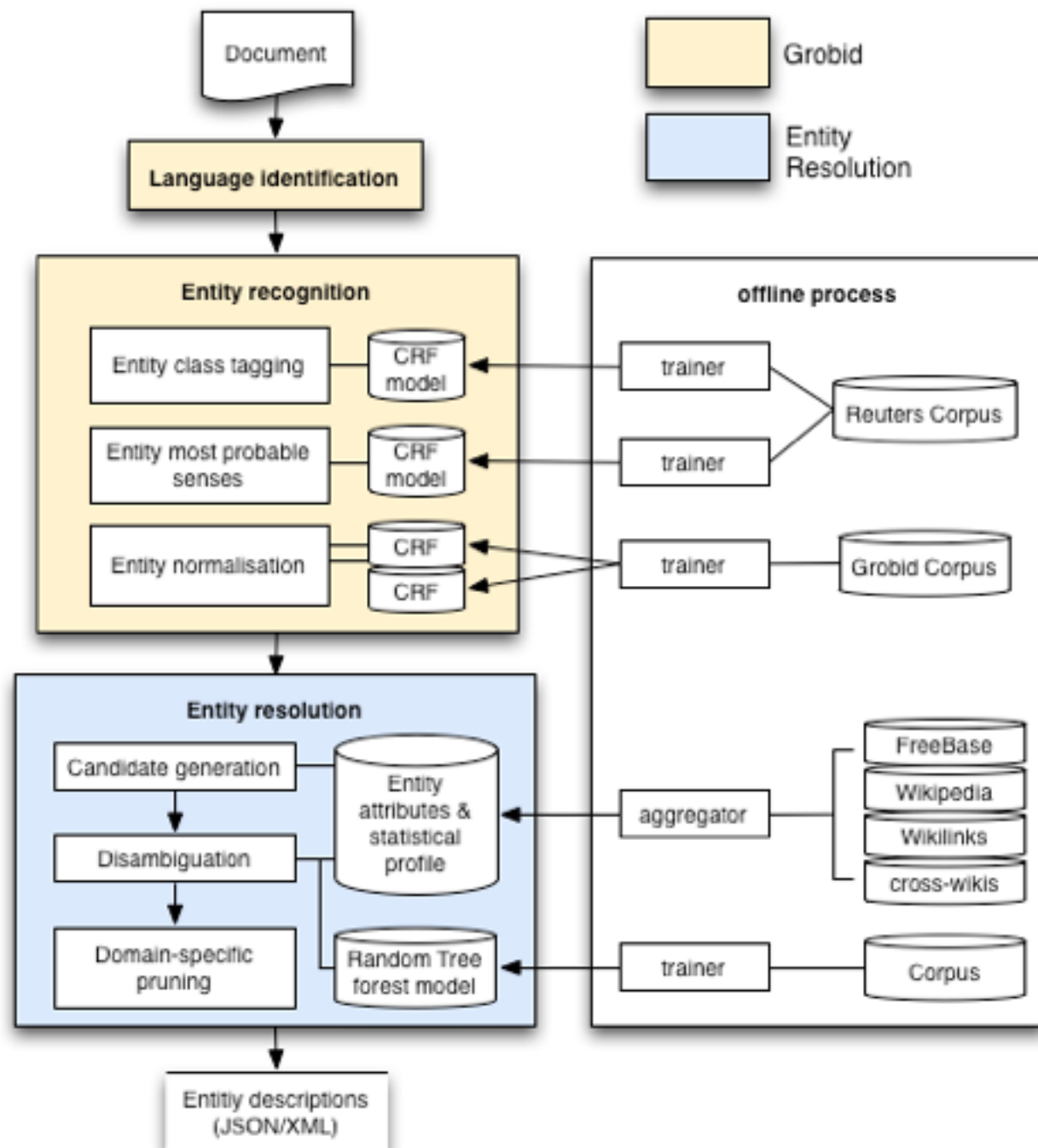
- Grobid: <https://github.com/kermitt2/grobid>
 - ➔ demo: <http://grobid.science-miner.com>
- anHALytics: <https://github.com/anHALytics>
 - ➔ demo: <http://traces1.saclay.inria.fr/anHALytics>
- (N)ERD: <https://github.com/kermitt2/grobid-ner> (partial!)
 - ➔ demo: <http://nerd.science-miner.com>
- GROBID-Quantity: <https://github.com/kermitt2/grobid-quantities>
 - ➔ demo: <http://quantity.science-miner.com>
- Keyterm extraction: not yet on GitHub
 - ➔ demo: <http://keyterm.science-miner.com>
- BEAST demonstrator: not yet on GitHub
 - ➔ demo : en chantier!

Travail du chantier d'usage ISTEEX

1. Annotations spécialisées en information scientifique et technique sur un grand échantillon - correspondant à 10 à 20% de l'ensemble du corpus ISTEEX
 - (N)ERD : entités connus / catégorisation
 - BIO-GROBID/BEAST : biotech
 - CHEMICAL-GROBID : chimie
 - grobid-quantities : mesures physiques
2. Encodage TEI stand-off des annotations
3. Démonstrateur exploitant ces annotations basé sur notre infrastructure anHALytics

Références

- Evans, J., Rzhetsky, A. "*Machine Science*". *Science*, vol. 329. no. 5990, pp. 399 - 400 (2010).
- Tony Hey, Stewart Tansley, and Kristin Tolle. "*The Fourth Paradigm: Data-Intensive Scientific Discovery*", (2009).
- Younger, S. M. "*Supercomputing and the Human Endeavor: The Coming Scientific Revolution in How We Use Machines to Help Us Think*". White paper, Los Alamos National Laboratory, (2001).



NERD

Features utilisées pour la désambiguisation générées pour chaque candidat

- commonness (probabilité du concept pour un terme)
- relatedness with local context
- relatedness with glocal context
- FreeBase type
- Inverse probability (probabilité du terme pour exprimer le concept)
- NER class
- NER sense
- *KL divergence* entre le texte de définition du candidat (par exemple article Wikipedia) et le contexte textuel de mention du terme

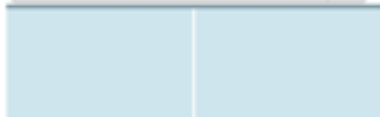
→ Approche **Machine Learning supervisée** : Les features sont utilisées pour entraîner un Random Tree Forest à l'aide de données d'entraînement existants (Yahoo! Webscope L24, AIDA/CoNNL-YAGO and ERD-50)

→ Approche **Machine Learning non supervisée** : entraînement complété par une sélection d'articles Wikipedia pour exploiter les ancres/liens vers articles



+ add new facet

normalizedDate



17 | 18

DD MM YYYY to DD MM YYYY ✓

Library

- Bridgeman Digital Images (6)
- NLS digital collections (2)

SubjectTerms

- German (8)
- fighter pilot (6)
- World War One (6)
- Photograph - 20th century (6)
- Great War (6)
- First World War (6)
- portrait (5)
- male (5)
- Armed Forces - Airforce (5)
- luftwaffe (4)

Language

Q red baron

Disamb./Expand

8 results - in 16 ms (server time)

The Red Baron (b/w photo).

Anonymous - Published date unknown

... Photographer false 20th false Bridgeman Digital Images WAR & MILITARY SCENES: 20TH CENTURY UNDETCOP German Photographer Publisher The Red...

... Baron (b/w photo) fighter pilot Armed Forces - Airforce Photograph - 20th century hero uniform aeroplane airplane cap male Air...



The Red Baron (b/w photo).

Anonymous - Published date unknown

... false 20th false Bridgeman Digital Images WAR & MILITARY SCENES: 20TH CENTURY UNDETCOP German Photographer Publisher The Red Baron...



Captain Baron VON RICHTHOFEN landing his FOKKER TRIPLANE (b/w photo).

Anonymous - Published date unknown

... travel Historical Events 1900 - present day World War One German World War I and World War II First World War Red Baron UNDETCOP ...

... & MILITARY SCENES: 20TH CENTURY UNDETCOP German Photographer Publisher Captain Baron von Richthofen landing his Fokker Triplane (b/w photo...



Author(s): German Photographer

Subject Terms: World War I (1914-18), fighter pilot, Photograph - 20th century, aeroplane, airplane, luftwaffe, Air, fighter plane, Great War, Air travel, Historical Events 1900 - present day, World War One, German, World War I and World War II, First World War, Red Baron

Copyright: undetermined copyright

FOKKER TRIPLANE

Type: artifact

Sub-type: undefined

conf: 0.656