Extraction, Exploitation and Evaluation of Document-Based Knowledge

Différentes approches généralistes du texte

Antoine Doucet
http://www.info.unicaen.fr/~doucet

Habilitation à Diriger des Recherches

Plan

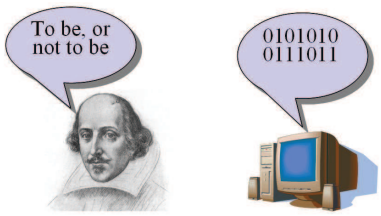
Introduction
Angle Séquentiel
Fouille de données textuelles, statistique appliquée
Application en RI multilingue
Angle Discursif
Veille Multilingue
Angle Structurel
RI Structurée
Clustering
Évaluation
Book Search Track
Conclusion

Enjeux de la description documentaire

- Le traitement documentaire requiert un grand nombre de comparaisons entre documents
  - Un moteur de recherche Internet doit par exemple décider quelles pages vont vous convenir parmi quelques milliards, étant donnée votre préférence, définie par la requête soumise
  - Une réponse en une seconde est considérée comme lente

Modéliser des documents

- Il faut donc pouvoir comparer les documents très efficacement
- Une réponse naturelle est d'opter pour des représentations de document simples



Pourquoi « générique et multilingue » ?

- Forte corrélation avec le passage à l'échelle
- La part de l'anglais sur Internet décroît
- De très nombreuses langues sont « peu dotées »
- Ce n'est pas un dogme :
  - Pose de jalons génériques
    - dont les performances peuvent souvent être améliorées par des méthodes spécifiques au corpus

Utilisation de la nature séquentielle de la donnée textuelle

- Travaux de thèse : *Advanced Document Description : a Sequential Approach*
  - Extraction, Sélection et Exploitation de séquences d'items (contexte applicatif textuel non restrictif)
- Pour le texte :
  - Méthodes entièrement généralistes (notamment multilingues)
  - Unités lexicales complexes
    - « cordon bleu » $\neq$ cordon + bleu
    - « cordon vert » ? « café puissant » ?
    - *red cell phone*

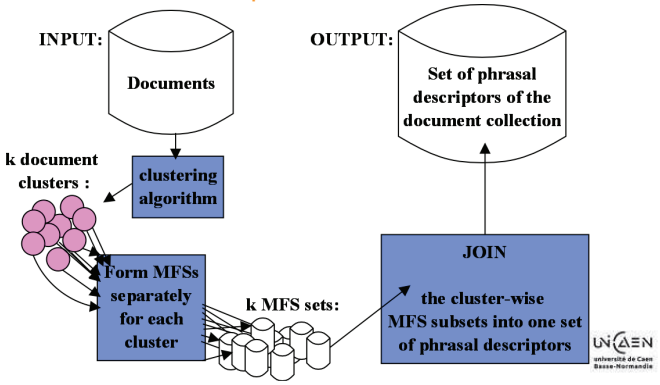
Extraction de motifs séquentiels discontinus I

- Séquentiels : la seule contrainte est l'ordre
  - dans l'état de l'art : contraintes sur la distance, la fréquence, la longueur, des motifs linguistiques…
- Discontinus : qu'importe la distance les séparant
- Séquences Fréquentes Maximales (MFS)
  - Maximales et Fréquentes
  - Description compacte
    - Une 10-séquence remplace $\binom{10}{2} = 45$ paires
    - Aucune contrainte de longueur
    - Distance illimitée entre les composants

Extraction de motifs séquentiels discontinus II

- On peut noter la similarité entre les deux fragments :
  - « ancient Président Clinton »
  - « l'ancient Président des États-Unis Bill Clinton »
- Méthode
  - Expansion par combinaison des 2- et 3-séquences fréquentes
  - Diviser pour régner

Extraction de motifs séquentiels discontinus III



Filtrage des séquences extraites I

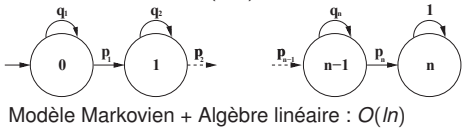
- Un ensemble de séquences descriptives est ainsi relié à chaque document
- Beaucoup sont très peu discriminantes : « *the be the a* », « *in the of* », « *barrel dollar* »…
- Probabilité d'occurrence d'une séquence discontinue
  - Expectative de fréquence documentaire
    - Test statistique permettant l'évaluation directe de l'intérêt des séquences (mesure de cohésion lexicale)
  - Réalisation d'un classement automatique
    - Indépendant du contexte applicatif (lexicographie, RI)
    - Combinant les séquences de différentes tailles

Filtrage des séquences extraites II

- Calcul de la probabilité d'occurrence d'une séquence discontinue $p(A_1 \to \cdots \to A_n, l)$
  - Approche naïve : $O(ln^{l-n})$
    - Disjonction par position initiale de succès
    - Soit E_i l'ensemble des documents contenant la n-séquence après exactement $(n+i)$ item :
      $$E_i = \{\bar{A_1}^{k_1} A_1 \bar{A_2}^{k_2} A_2 \ldots \bar{A_n}^{k_n} A_n W^{l-n-1} \mid \sum_{i=1}^{n} k_i = i\}$$
  - Or les (E_k) sont disjoints, donc :
    $$p(A_1 \to \cdots \to A_n, l) = \prod_{i=1}^{n} p_i \sum_{i_n=0}^{l-n} \cdots \sum_{i_1=0}^{l-n-(i_n+\cdots+i_2)} q_1^{i_1} q_2^{i_2} \ldots q_n^{i_n}$$

Filtrage des séquences extraites III

- Calcul de la probabilité d'occurrence d'une séquence discontinue $p(A_1 \to \cdots \to A_n, l)$
  - Approche naïve : $O(ln^{l-n})$
  - Modèle Markovien : $O(ln^3)$



  - Modèle Markovien + Algèbre linéaire : $O(ln)$
    - Propriétés spécifiques à la matrice de transition, notamment Jordanisation à coût algorithmique réduit

Application en RI multilingue I

- Appariements de phrases avec des paramètres variables
  - Inversion des composants
  - Pénalité de distance
  - Distance d'appariements maximales
- Corpus spécialisés et non-spécialisés
  - Articles scientifiques
  - Dépêches d'agence de presse
- Quatre langues de familles et graphies distinctes
  - Chinois, Coréen, Japonais, Anglais
- (Re)découverte de propriétés linguistiques

Application en RI multilingue II

- Cette mesure de similarité multi-mots est plus profitable que l'ajout de paires adjacentes à l'ensemble des descripteurs

	WVSM	SEQ-Big	SEQ-Adv
NTCIR-CH	(0.1705)	0.1327*	**0.1885***
NTCIR-JP	(0.2151)	0.1480*	**0.2246***
NTCIR-KR	(**0.1707**)	0.1049*	**0.1499**
NTCIR-EN	(0.2555)	**0.2692**	0.2208*
INEX(EN)	(0.04193)	**0.04935***	0.04370

Contributions

- Extraction de séquences, sous contraintes réduites
  - Descripteurs compacts
  - Passage à l'échelle
- Sélection de séquences
  - Méthode de calcul de la probabilité d'occurrence d'une séquence discontinue en $O(ln)$
- Application en RI
  - Mesure de similarité multi-mots
- Ressources endogènes
- Techniques généralisables à tout type de donnée séquentielle

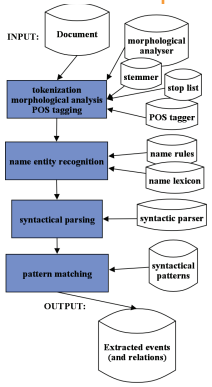
Perspectives

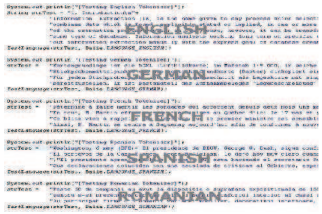
- Alignement de paraphrases
  - Extraction de synonymes
- Extraction de terminologie
  - Construction de *thesauri* multilingues
- Application à d'autres types de données séquentielles
  - Autres données linguistiques
  - Bioinformatique
  - Panier de la ménagère

Veille épidémiologique

- Objectifs
  - Détecter, dans un flux de dépêches, les évènements sanitaires et les localiser dès leur survenue
  - Public : épidémiologistes (OMS, INVS, CE)
- « Dès la survenue » $\Longrightarrow$ souvent dans la langue locale
  - L'impératif de réactivité nécessite le traitement d'un maximum de langues

L'approche classique multilingue (e.g., PULS)

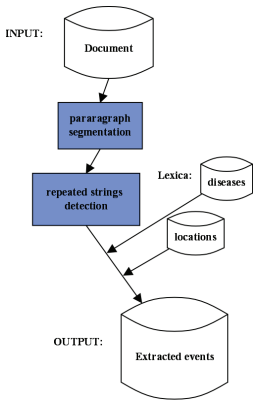




n langues = *n* analyseurs à grain phrase

L'approche MultiPULS

- Approche simplifiée au grain texte
- avec ressources minimales pour le filtrage des évènements



Application d'un modèle rhétorique

- Les positions sont exploitées
- Les répétitions sont exploitées
  - Au grain caractère
  - Par extraction de chaînes répétées maximales
- Les paires significatives sont extraites
  - Maladie - Lieu
  - Maladie - Nombre de cas

Exemple : texte brut

WHO checks smallpox reports in Uganda

The World Health Organisation said today it was investigating reports of suspected cases of the previously eradicated disease smallpox in eastern Uganda.

Smallpox is an acute contagious disease and was one of the world's most feared sicknesses until it was officially declared eradicated worldwide in 1979. "WHO takes any report of smallpox seriously" Gregory Hartl, a spokesman for the Geneva-based United Nations health agency, told Reuters via email. "WHO is aware of the reports coming out of Uganda and is taking all the necessary measures to investigate and verify."[...]

Exemple : répétitions

WHO checks **smallpox reports** in **Uganda**

The World Health Organisation said today it was **investigat**ing **reports** of suspected cases **of the** previously **eradicated disease smallpox** in eastern **Uganda**.

Smallpox is an acute contagious **disease** and was one of **the world**'s most feared sicknesses until it was officially declared **eradicated** worldwide in 1979. "**WHO** takes any **report** of **smallpox** seriously" Gregory Hartl, a spokesman for the Geneva-based United Nations **health** agency, told Reuters via email. "**WHO** is aware **of the reports** coming out of **Uganda** and is taking all the necessary measures to **investigat**e and verify."[...]

Exemple : répétitions + filtrage lexical maladie/lieu

WHO checks **smallpox reports** in **Uganda**

The World Health Organisation said today it was **investigat**ing **reports** of suspected cases **of the** previously **eradicated disease smallpox** in eastern **Uganda**.

Smallpox is an acute contagious **disease** and was one of **the world**'s most feared sicknesses until it was officially declared **eradicated** worldwide in 1979. "**WHO** takes any **report** of **smallpox** seriously" Gregory Hartl, a spokesman for the Geneva-based United Nations **health** agency, told Reuters via email. "**WHO** is aware **of the reports** coming out of **Uganda** and is taking all the necessary measures to **investigat**e and verify."[...]

Un exemple en polonais

Disease Place Number of Cases show the keywords found by daniel

Document language : pl

Czarne legginsy niebezpieczne dla zdrowia

Tajlandzki rząd ostrzega kobiety przed noszeniem czarnych legginsów, gdyż ciemne kolory przyciągają komary, przenoszące dengę. Choroba ta w tym roku zabiła już w Tajlandii 43 osoby - podała agencja Associated Press.

Martwi nas sposób ubierania się młodych ludzi - poinformowała w wydanym w niedzielę oświadczeniu wiceminister zdrowia Pansiri Kulanartsiri. - Sugerujemy, by unikali noszenia czarnych legginsów, a także innych ubrań w tym kolorze, by nie przyciągać komarów.

Noście grube ubrania, na przykład jeansy - radziła wiceminister.

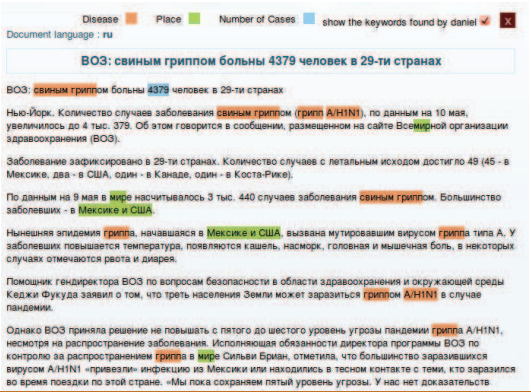
W tym roku w Tajlandii odnotowano ponad 45 tys. przypadków dengi, czyli o 40% więcej niż w ubiegłym roku. Na chorobę tę do końca lipca zmarły aż 43 osoby; 26 z nich było w wieku od 10 do 25 lat.

Przewiduje się, że sytuacja pogorszy się podczas pory deszczowej, która rozpoczęła się w czerwcu i potrwa do września. W tym okresie stojąca woda i bagna stają się wylęgarnią komarów.

Denga, która występuje głównie w wielkich miastach, jest ostrą chorobą zakaźną, wywoływaną przez wirusy przenoszone przez komary Aedes aegypti oraz Aedes albopictus. Do jej symptomów należą m.in. gorączka, bóle mięśniowe i brzucha, wysypka czy obrzęk węzłów chłonnych. Jej najpoważniejsza forma powoduje krwotoki wewnętrzne, powiększenie wątroby oraz niewydolność układu krążenia.

Nie istnieje lekarstwo na dengę, a według Światowej Organizacji Zdrowia (WHO) szczepionka będzie dostępna dopiero za parę lat.

Un exemple en russe



Un exemple en arabe

Résultats

- Résultats proches de l'état de l'art...

Langue	Taille Corpus	Rappel	Précision
Français	1954	92%	84%
Anglais	540	97%	84%
Russe	400	88%	85%
Polonais	439	85%	73%
Chinois	100	92%	85%

- ... pour la minorité de langues qui y sont traitées

Langue	Anglais	Français	Russe	Polonais	Chinois
MultiPULS	84%	84%	85%	73%	85%
Biocaster	93%	n/a	n/a	n/a	n/a

Bilan

- Contexte : Thèse de Gaël Lejeune
  - Encadrante principale : Nadine Lucas
  - Collaboration avec l'université d'Helsinki (PICS Multipuls 2009–2011)
- Une chaîne de traitement alingue complète
- Un faible coût en ressources
  - lexique d'environ 500 termes par langue
  - 100 fois inférieur à l'état de l'art
- Un temps de traitement satisfaisant
  - 1000 documents par minute
  - 10 fois plus rapide que PULS
- Une plateforme d'annotation et de validation en ligne

Perspective : détection de fraîcheur

- Vers une veille indépendante du domaine
  - Pas de ressources → indépendance du domaine
  - Le choix éventuel d'un domaine se fait *a posteriori*
- Détection d'associations lexicales fortement divergentes
- Thèse d'Oskar Gross
  - Encadrant principal : Hannu Toivonen (U. Helsinki)
  - Projet d'ANR Sucrett - *SUpporting CREativity from TexT*
  - Semestre à Helsinki en cours

Exploitation de la structure des documents

- Utiliser la structure logique des documents (e.g., sections, sous-sections paragraphes), stylistique (e.g., gras, italique)
- De nouvelles possibilités :
  - Répondre à un besoin d'information par des fragments de documents précis, en tirant profit de leur structure logique
  - Affiner les réponses par granularité structurelle
- La notion de granularité s'ajoute à celle de pertinence
- Deux contributions :
  - RI structurée
  - Partitionnement de documents

Système EXTIRP

EXacT coverage IR based on static Passage clusters

1. Détection des unités minimales de recherche
2. Calcul de leur pertinence
3. Propagation verticale dans l'arbre XML du document
4. Sélection du grain adéquat

Metric	@1	@5	@10	@100	@1500
ncXG (strict)	8	12	15	34	27
ncXG (generalised)	13	26	29	32	22
inex_eval (strict)	1	2	11	26	22
inex_eval (generalised)	11	19	25	31	24

[rangs INEX 2005 (sur 44)]

Système EXTIRP : amélioration

- Notion de « *textitude* »d'un marqueur XML
  - *T/E (Text node/Element node) ratio*
  - Différencie les marqueurs XML de structuration de ceux de mise en forme
  - Duplication en ligne des textes mis en forme

```
...kernel trick has been applied to several algorithms in
<link>machine learning</link> <link>machine learning</link> and
<link>statistics</link> <link>statistics</link>, including...
```

- Utile notamment pour l'extraction de séquences !

	Original	all Dupl.	Hybrid
iP 0.01 (107 topics)	0.3319	0.3773	**0.3815**
MAiP (107 topics)	0.0912	0.1024	**0.1036**

[résultats INEX 2007 – SIGIR 2008]

Partitionnement de documents XML

- But : prendre en compte la structure dans le processus de classification
  - Génération automatique de DTDs
  - *Cluster hypothesis (a priori)*
- Méthode de clustering en 2 étapes
  1. Clustering basé sur des descripteurs structurels exclusivement : Détection des « outliers » structurels
  2. Pour les clusters dont la similarité interne est inférieure à un seuil : second clustering basé sur des descripteurs textuels classiques

Features	Text	Tags	Tags + Text	Tags → Text
Entropy	0.633	0.798	0.678	**0.630**
Purity	0.379	0.228	0.372	**0.394**
Clustering Time	754s.	11s.	837s.	**11+742s.**

Contributions

- Généralisme : aucune connaissance préalable requise
  - fonctionne sans DTD ou *Schema* XML
- RI structurée – Projet *EXTIRP*
  - Mené à l'université d'Helsinki 2002–2003 (6 personnes)
  - Participation annuelle à INEX jusqu'à 2009
- Classification non supervisée
  - Système vainqueur *INEX Mining track* 2006 (Wikipedia)
- Exploitation de la *textitude* des nœuds XML
  - Thèse de Miro Lehtonen (2006)

Perspectives

- RI structurée
  - RI dans des bibliothèques numériques/numérisées
- Classification non supervisée
  - Classification de fragments de documents
  - Résumé multi-documents

Travaux en méthodologie d'évaluation

- Objectif
  - Évaluation de tâches sur des collections d'ouvrages numérisés
  - Deux contributions :
    - Évaluation de la performance des SRI
    - Évaluation de la performance en extraction de structure

Contexte

- Contexte : *INEX Book Search Track*
- Un cadre d'application clé
  - Accès à des collections d'ouvrage en ligne
    - Trouver un livre
    - Trouver de l'information dans des livres
  - Numérisation massive
    - *Google Books*
    - Bibliothèques nationales
  - Apporter le livre ancien aux *e-books*, liseuses électroniques


Numérisation

1. Livre (papier)
2. Photos des pages (.jpg)
3. OCR (.xml)
   - Bruit
   - Nécessite des standards

Kirtas book scanner

Difficultés spécifiques

- Grandes collections de grands documents
- Structure physique plutôt que logique
- Absence de références croisées
- Bruit (livres anciens ← OCR)

Évaluer la performance des SRI

- Défi principal : collection des annotations
  - Développement d'une plateforme
    - `http://www.booksearch.co.uk`
  - Compétition d'annotateurs (jeu)
  - *Crowdsourcing*

Évaluer la performance de l'extraction de structure

- Aucune méthodologie adéquate existante
  - Définition d'un cadre d'évaluation complet
  - Définition de mesures d'évaluation
  - Construction de la vérité de terrain
    - Processus d'annotation collaborative avec plateforme Java distribuée

Contributions

- *INEX Book Search track*
  - Avec Gabriella Kazai (*Microsoft Research Cambridge*)
  - Atelier de travail annuel pour les participants
- Mise en place de méthodologies d'évaluation :
  - RI dans des collections d'ouvrages
    - Initié en 2007
    - Tâche principale d'INEX depuis 2011
    - Intégré à CLEF depuis 2012
  - Extraction de structure (ouvrages numérisés)
    - Initié à INEX 2008
    - Compétition ICDAR depuis 2009

Perspectives

- RI dans des collections d'ouvrages
  - RI sociale
  - Exploitation de résumés et de recommandations
- Extraction de structure (ouvrages numérisés)
  - *Crowdsourcing*
  - Évaluation indirecte

Bilan

PIPELINE DE TRAITEMENT	APPROCHE DU MATÉRIAU TEXTE
Extraction	Séquentielle
Exploitation	Discursive
Évaluation	Structurelle

CONSTANTE : GÉNÉRICITÉ, NOTAMMENT MULTILINGUISME

Perspectives

- Détection de fraîcheur
  - Détection d'associations lexicales fortement divergentes
  - Généralisation de la veille multilingue
    - Le choix éventuel d'un domaine se fait *a posteriori*
- RI contextuelle structurée (personnalisée)
  - RI structurée sur des collections d'ouvrages
  - Suggestions adaptées au contexte
- Recherche temporelle d'images
  - Datation automatique de photographies
    - corrélation entre contenu des images, date de prise de vue
    - environnement textuel et journaux de requêtes