

Contextualisation automatique de Tweets à partir de Wikipédia

Romain Deveaud¹ – Florian Boudin²

¹ LIA – Université d'Avignon

² LINA – Université de Nantes

Introduction

- l'accès à l'information
- smartphones

element

, amis...)

MAR 29 LinkedIn today The day's top news, tailored for you



La cyberguerre
ecrans.fr · Il paraît depuis plusieurs jours qu'elle a réussi à...

Like · Comment · Share · Save · 17h

Add a comment

1 patron sur

Tracy Chou @triketora 3h
Game Developers Conference Features Sexy Dancing Ladies, Because Tech Is Super Welcoming to Women prsm.tc/bx30XI
Retweeted by Mounia Lalmas
[View summary](#)

Claudia Hauff @CharlotteHase 57m
Used yesterday to scan through all the #ECIR2013 papers. Definitely worth the time!
Expand

YahooJobs @YahooJobs 1h
We're excited to announce we're expanding in Dublin. Check out our 200 new jobs in the Sales Ops Centre. bit.ly/11J2uIQ
Expand

Eric Hennekam @EricHennekam 2h
In de archief- en personenzoeker archieffzoeker.nl een hele serie nieuwe zoekwoorden verstoppt bv tracking en stamboom
Retweeted by Arjen P. de Vries
Expand

Tweets 4 Science @tweets4sci 10h
Help double the size of our repository, get a friend to donate their tweets to science today! tweetsforscience.org #tweets4science
Retweeted by Guillaume Cabanac
Expand

Eric Charton @ericcharton 7h
Enseignants et Wikipédia: faites ce que je dis, pas ce que je fais zdnnet.fr/39787878 via @zdnnetfr
Expand

inseph raisinger @insephraisinger 8h

Personalize this!
Tools to make Google News yours

Drill

ed Friday that

Related
North Korea »
Korean War »
warning to

Personalize Google News

Recent

to settle
retoric and a
y drills with

North Korea puts rockets on standby to "settle accounts" with US amid fears ...
CBS News - 17 minutes ago

Ousted Central African Republic leader Bozize seeks exile in W. African nation ...
Fox News - 7 minutes ago

Cyprus banks open for business for second day
Fox News - 7 minutes ago

Weather for Avignon

Today	Sat	Sun	Mon
17° 6°	16° 5°	14° 2°	14° 6°

The Weather Channel - Weather Underground - 2couWeather

Introduction

-  Twitter, un des vecteurs de diffusion d'information
 - messages courts, limités à 140 caractères
 - souvent des pointeurs vers des pages externes
- comment savoir de quoi parle un Tweet sans cliquer sur les liens?
 - contexte thématique

Introduction

- Twitter est une mine d'information en temps réel (mais aussi **beaucoup de bruit**)
 - les Tweets comme **réponses** à une requête? (TREC Microblog track)
 - encore faut-il les comprendre...
- représentation de ce contexte par un ensemble de phrases
 - **résumé** du contexte thématique

Introduction



iTunes TV @iTunesTV

13 Feb 12

What are your favorite shocking moments of [@WalkingDead_AMC](#) so far? [@AMC_TV](#) [#TheWalkingDead](#) tw.itunes.com/JLg

Expand

- AMC chaîne TV américaine
- The Walking Dead : série (science-fiction)
- moments choquants de cette série...

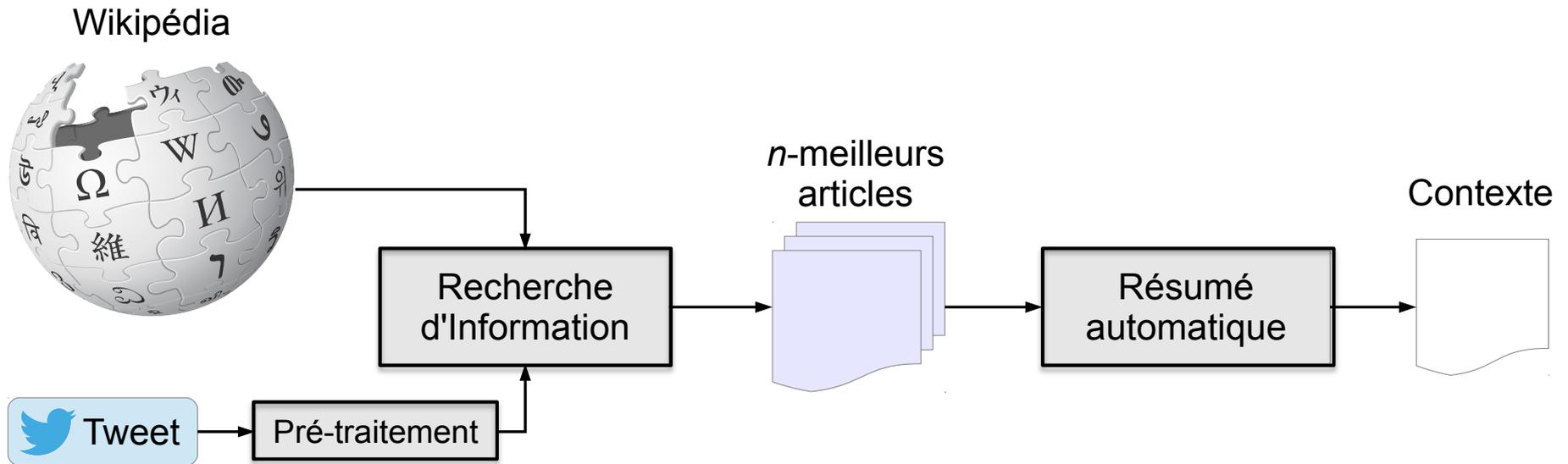
Introduction

- problématique actuelle, mettant en œuvre différents champs de recherche
 - recherche et extraction d'information, résumé automatique
- extraction de phrases à partir de Wikipédia
 - dans une limite de 500 mots
 - scénario mobile où un utilisateur lit sur son smartphone
 - INEX Tweet Contextualization



Introduction

- une vision partielle de la contextualisation
 - estimation de la **signification** globale d'un Tweet
 - réduction, jusqu'à n'avoir que les informations liées les plus représentatives



Pré-traitements



iTunes TV @iTunesTV

What are your favorite shocking moments of [@WalkingDead_AMC](#) so far? [@AMC_TV](#) [#TheWalkingDead](#) [tw.itunes.com/JLg](#)

13 Feb 12

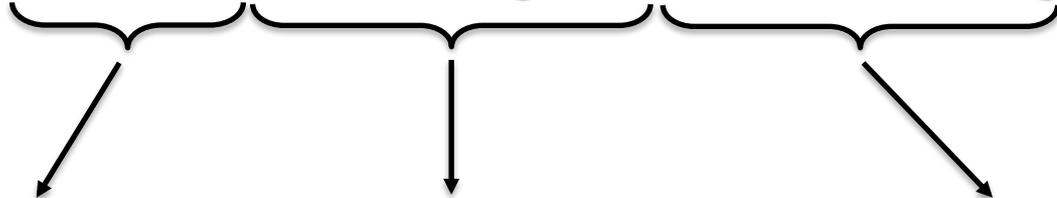
Expand

mention

hashtag

URL (non autorisée)

mention



Pré-traitements



iTunes TV @iTunesTV

13 Feb 12

~~What are your favorite shocking moments of @WalkingDead_AMC~~
~~so far? @AMC_TV #TheWalkingDead tw.itunes.com/JLg~~

Expand

$H_{\mathcal{T}}$ = “the walking dead”

\mathcal{T} = “favorite shocking moments walking dead”

} <title>

} <description>

parallèle avec les *topics* TREC

Recherche d'articles Wikipédia

- approche par modèle de langue
 - “Tweet likelihood”

$$P(\mathcal{T}|\theta_D) = \prod_{t \in \mathcal{T}} f_T(t, D)$$

- lissage standard de Dirichlet

$$f_T(t, D) = \frac{c(t, D) + \mu \cdot P(t|\mathcal{C})}{|D| + \mu}$$

Recherche d'articles Wikipédia

- l'adjacence de deux mots dans un Tweet peut être importante
 - Markov Random Fields (MRF) [Metzler & Croft, SIGIR'05]

$$s_{MRF}(\mathcal{T}, D) = \lambda_T \prod_{t \in \mathcal{T}} f_T(t, D) + \left. \begin{array}{l} \lambda_O \prod_{i=1}^{|Q|-1} f_O(t_i, t_{i+1}, D) + \\ \lambda_U \prod_{i=1}^{|Q|-1} f_U(t_i, t_{i+1}, D) \end{array} \right\} \begin{array}{l} \text{"Tweet likelihood", unigrammes} \\ \text{bigrammes} \\ \text{bigrammes, au sein d'une fenêtre} \\ \text{non ordonnée} \end{array}$$

$$\lambda_T = 0,85, \lambda_O = 0,10 \text{ et } \lambda_U = 0,05$$

Intégration de *hashtags*

- étiquettes définies **manuellement**
 - simplification du Tweet
 - expression des informations les plus importantes

$$s(\mathcal{T}, H_{\mathcal{T}}, D) = \alpha \underbrace{s_{MRF}(H_{\mathcal{T}}, D)}_{\text{uniquement les } \textit{hashtags}} + (1 - \alpha) \underbrace{s_{MRF}(\mathcal{T}, D)}_{\text{tout le texte du Tweet}}$$

« U Just Heard “Hard To Believe” by [@andydavis](#) on the [@mtv](#) Teen Mom 2 Finale go 2 <http://t.co/iwb2JuL8> for info #ihearditonMTV »

Intégration de *hashtags*

- importance contextuelle des *hashtags*
 - estimation de leur pouvoir discriminant
 - score de clarté [Cronen-Townsend, HLT'02]

$$\alpha = \sum_{w \in V} P(w|H_{\mathcal{T}}) \log \frac{P(w|H_{\mathcal{T}})}{P(w|\mathcal{C})}$$

- modèle de langue des *hashtags* estimé par retour de pertinence simulé

Formation du contexte

- entrée : un ensemble d'articles Wikipédia liés à un Tweet
 - on choisit empiriquement les 5 mieux classés
 - résumé multi-document
- attribuer un score à chaque phrase
 - utilisation de différentes caractéristiques
 - les mieux classées constituent le contexte (dans une limite de 500 mots)

Formation du contexte

- quatre catégories de caractéristiques
 - importance de la phrase candidate dans le **document dont elle provient** (1) [Mihalcea,ACL'04]
 - pertinence de la phrase par rapport au **Tweet** (4)
 - pertinence de la phrase par rapport au **contenu d'une page web** dont l'URL est dans le Tweet (4)
 - pertinence du **document** dont provient la phrase (1)
- score final d'une phrase : combinaison linéaire

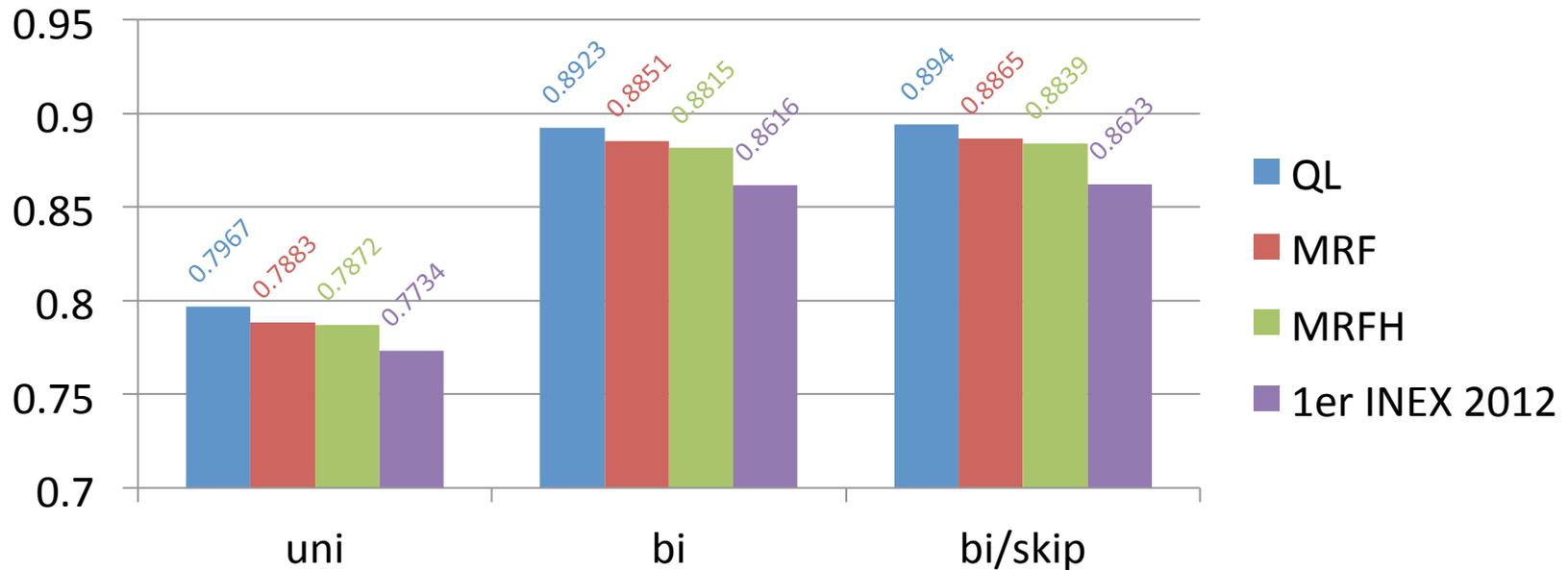
Evaluation

- collection de test de la tâche Tweet Contextualization d'INEX
 - version de Wikipédia : Novembre 2011
 - 1126 Tweets à contextualiser (63 évalués)
- indexation de Wikipédia avec Indri
 - suppression des mots outils
 - racinisation légère de Krovetz

Evaluation

- 3 approches évaluées
- QL : “Tweet Likelihood”
 - pas de *hashtags*, pas de dépendances entre les mots
- MRF : Markov Random Fields, sans *hashtags*
- MRFH : MRF avec *hashtags*, et influence calculée automatiquement

Evaluation



- les plus petits scores sont les meilleurs
- aucune différence significative entre les approches
 - peu dépendantes de la sélection des documents

Evaluation

- MRFH sensiblement meilleure que MRF
 - mais seulement 23% de Tweets possèdent des *hashtags*
- biais dans l(a métrique d)'évaluation?
 - plus d'informations dans la présentation suivante?

Evaluation

- *pooling* et jugements des 10 premières phrases
 - lisibilité et cohérence : les phrases + informatives ne se trouvent pas forcément en premier
- la métrique d'évaluation ne considère que des phrases jugées pertinentes
 - pas de références de phrases non-pertinentes
 - augmenter la diversité des phrases peut augmenter les scores

Evaluation

- régression logistique sur les caractéristiques

Caractéristique	Nom	Valeur	Significativité
<i>c1</i>	TextRank	8.996	$p < 2^{-16}$
<i>c2</i>	Recouvrement Tweet	2.496	$p = 2.38^{-6}$
<i>c3</i>	Cosine Tweet	5.849	$p = 4^{-15}$
<i>c4</i>	Recouvrement <i>hashtags</i>	-2.051	$p = 0.1368$
<i>c5</i>	Cosine <i>hashtags</i>	0.671	$p = 0.3074$
<i>c6</i>	Recouvrement titre URL	1.373	$p = 0.2719$
<i>c7</i>	Cosine titre URL	0.788	$p = 0.6287$
<i>c8</i>	Recouvrement page URL	0.543	$p = 0.4337$
<i>c9</i>	Cosine page URL	10.374	$p = 0.0195$
<i>c10</i>	Score document	0.782	$p < 2^{-16}$

Conclusion

- les *hashtags* d'un Tweet semblent apporter des informations contextuelles (sous réserve...)
- approche favorisant les Tweets factuels, *news*

Conclusion

- une phrase contextuellement importante...
 - ... apparaît dans un document bien classé
 - ... contient les mêmes mots que le Tweet
 - ... fait partie des phrases centrales du document dont elle est extraite
- reconnaissance d'entités nommées
 - *entity linking* avec Wikipédia
- détection des concepts implicites exprimés

merci de votre attention