Contextualisation automatique de Tweets à partir de Wikipédia

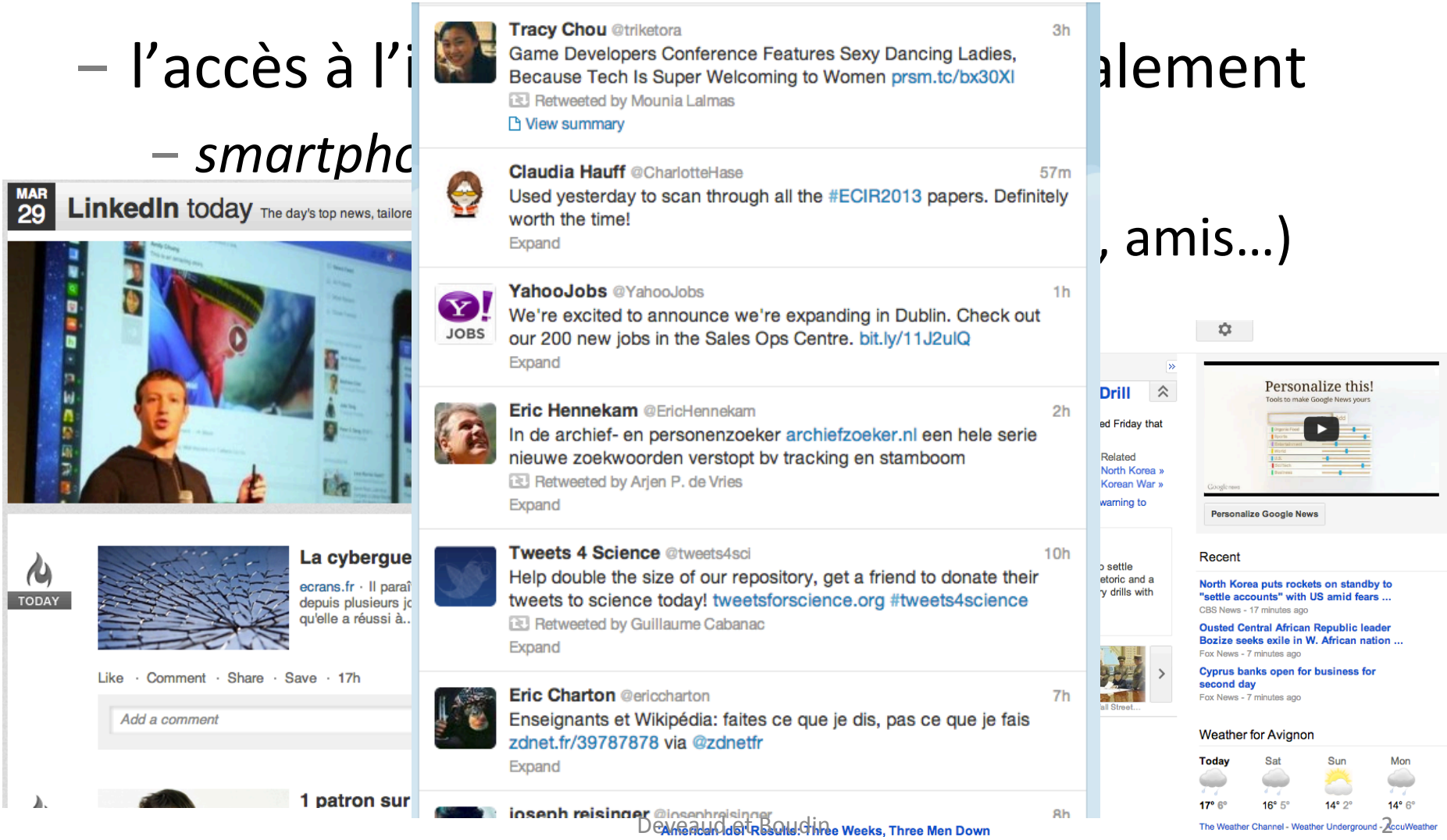
Romain Deveaud[1] – Florian Boudin[2]

[1] LIA – Université d'Avignon

[2] LINA –Université de Nantes

Introduction

- l'accès à l'i… alement
  - *smartpho…*



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

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

Introduction

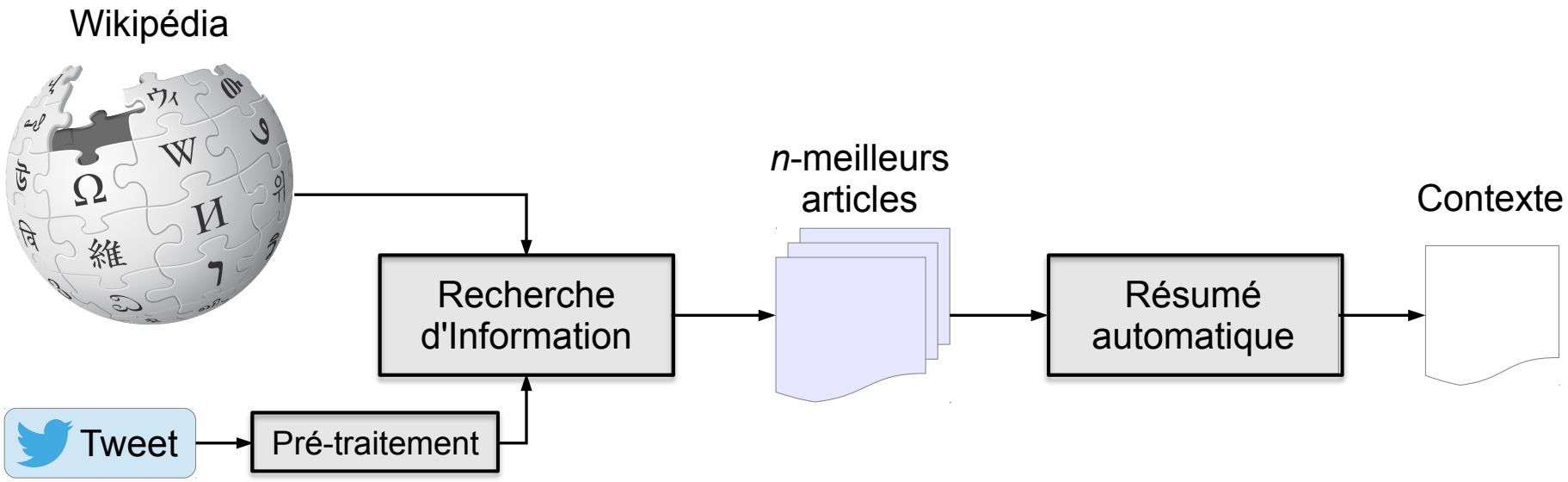
- problématique actuelle, mettant en œuvre différents champs de recherche
  - recherche et extraction d'information, résumé automatique

- 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

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” } **<title>**

$\mathcal{T}$ = “favorite shocking moments walking dead” } **<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]

$$s_{MRF}(\mathcal{T}, D) = \lambda_T \prod_{t \in \mathcal{T}} f_T(t, D) + \quad \Big\} \text{ "Tweet likelihood", unigrammes}$$

$$\lambda_O \prod_{i=1}^{|Q|-1} f_O(t_i, t_{i+1}, D) + \quad \Big\} \text{ bigrammes}$$

$$\lambda_U \prod_{i=1}^{|Q|-1} f_U(t_i, t_{i+1}, D) \quad \Big\} \text{ bigrammes, au sein d'une fenêtre non ordonnée}$$

$\lambda_T = 0,85$, $\lambda_O = 0,10$ 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 #iheardit onMTV »

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

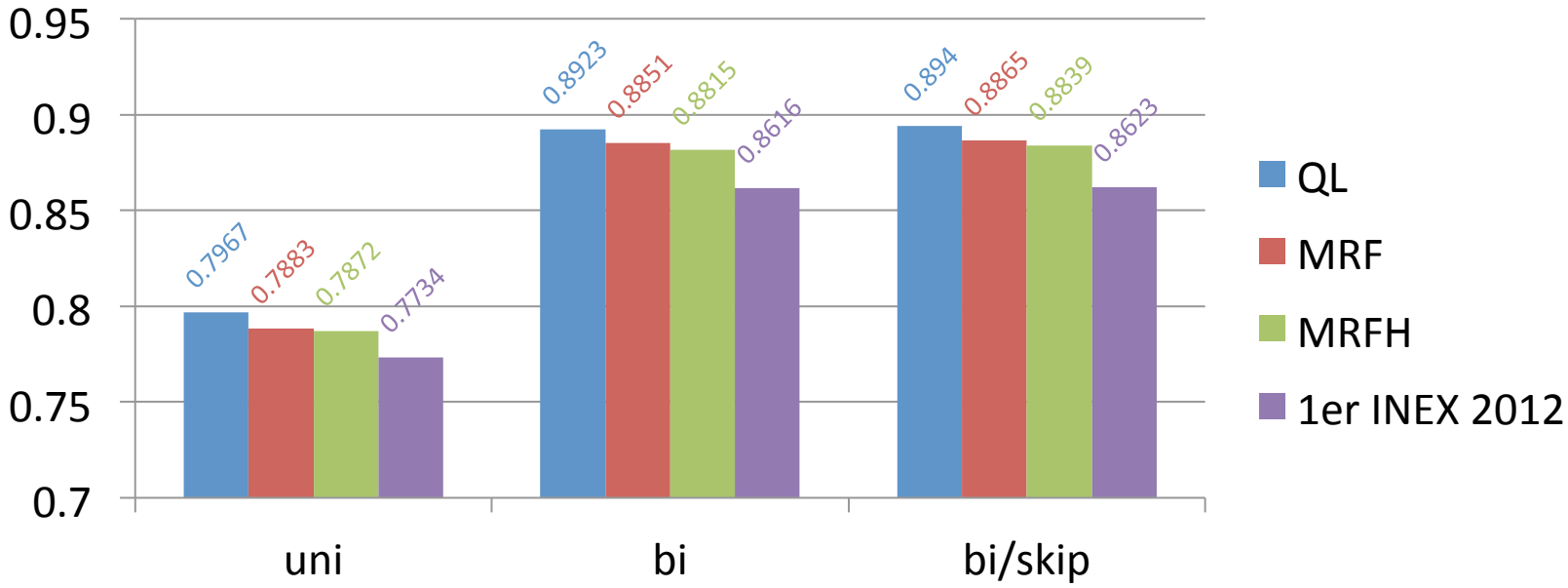
- collection de test de la tâche Tweet Contextualization d'INEX
  - version de Wikipédia : Novembre 2011
  - 1126 Tweets à contextualiser (63 évalués)

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