Traitement Automatique des Langues & Recherche d'Information

Vincent Claveau

Journée ARIA-AFIA 12/12/2017

Avant-propos

Intelligence Artificielle

- but : modeler l'intelligence humaine?
  copier de capacités humaines ou autres?
- moyens : modèle logique, statistique, apprentissage, DNN ?

Le TAL comme partie de l'IA

- test de Turing ⇒ chatbots

Avant

Intelligen

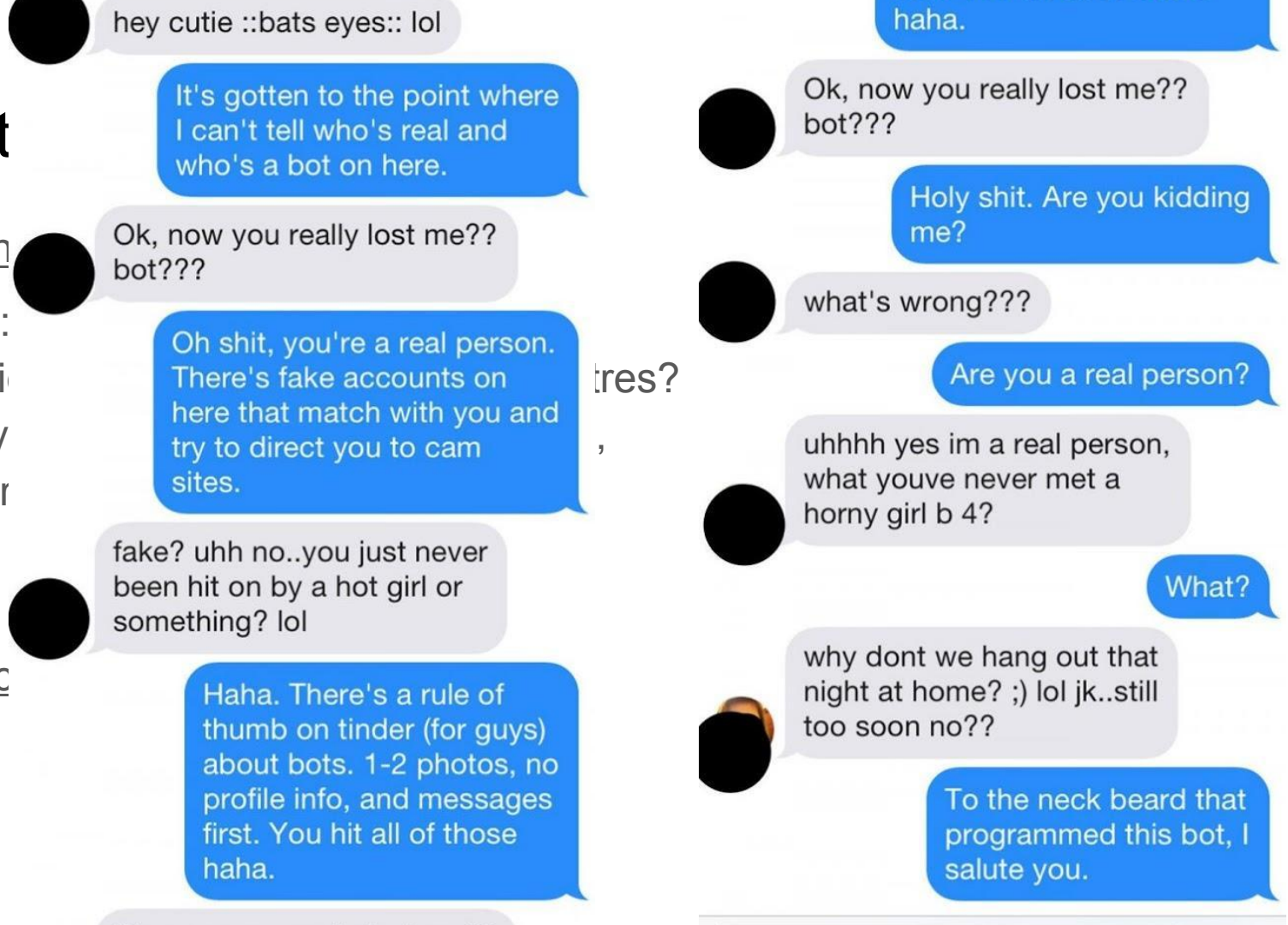
- but :
  copi
- moy
  appr

Le TAL c

- test

tres?

,



Avant-propos

Information Retrieval

Niveaux d'analyse

Phonétique
- Sons -> mots

Morphologie
- Formation des mots

Syntaxe
- Formation d'énoncés/phrases

Sémantique
- Sens des mots/phrases

Pragmatique
- Contexte, connaissance du monde

Déroulé

RI pour le TAL

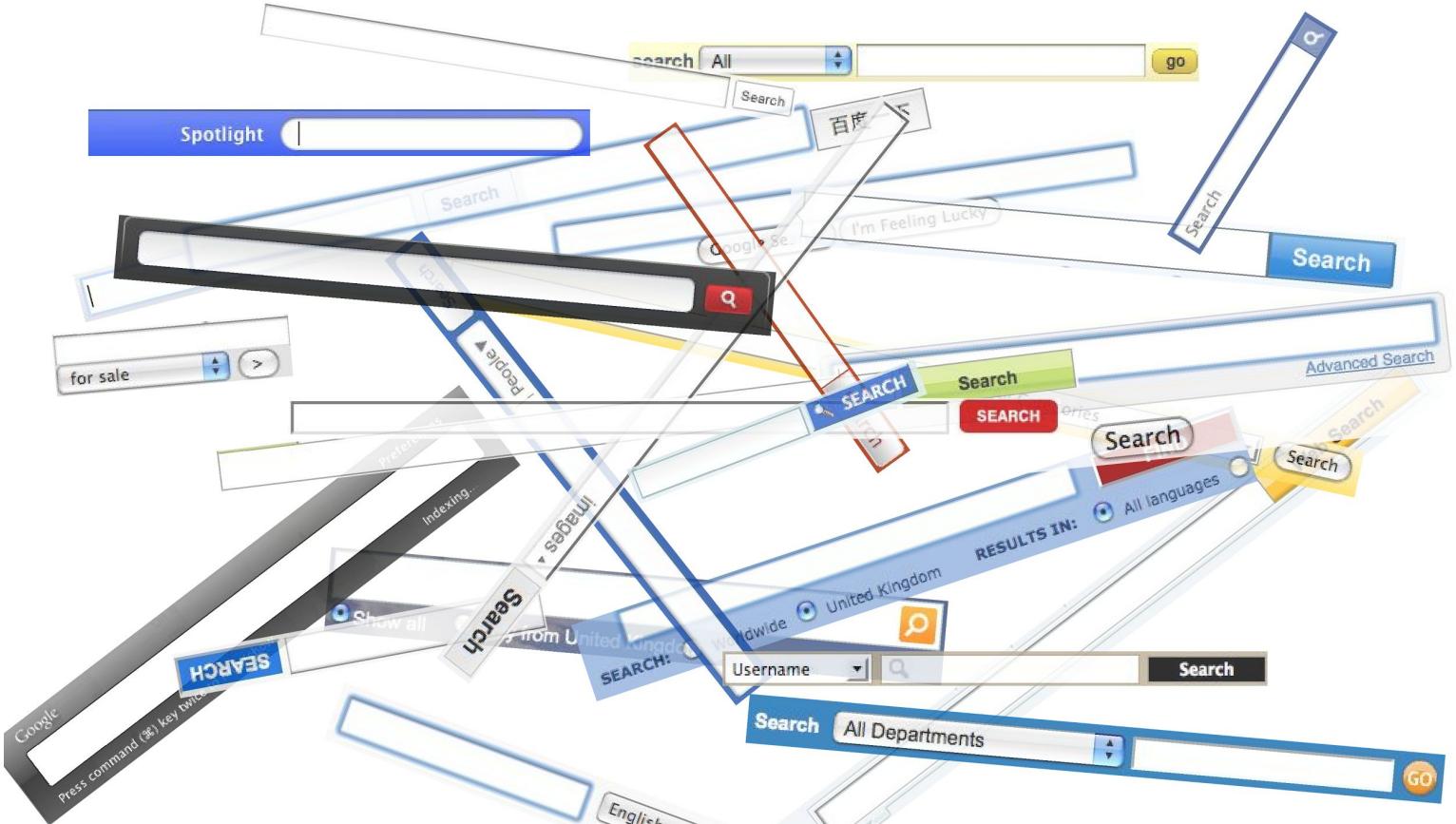
- Zoom sur deux applications

TAL pour la RI

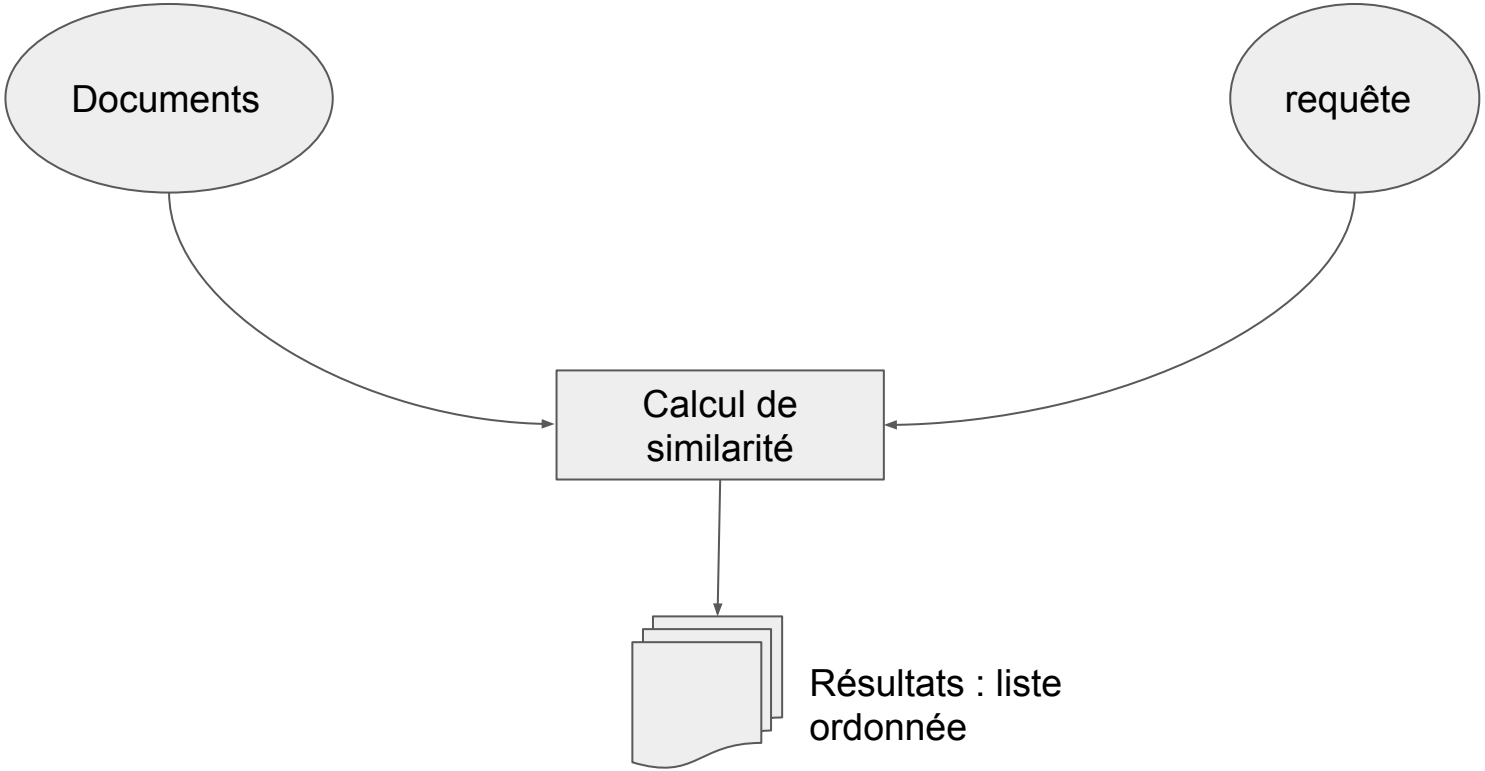
- Zoom sur deux applications

RI pour le TAL

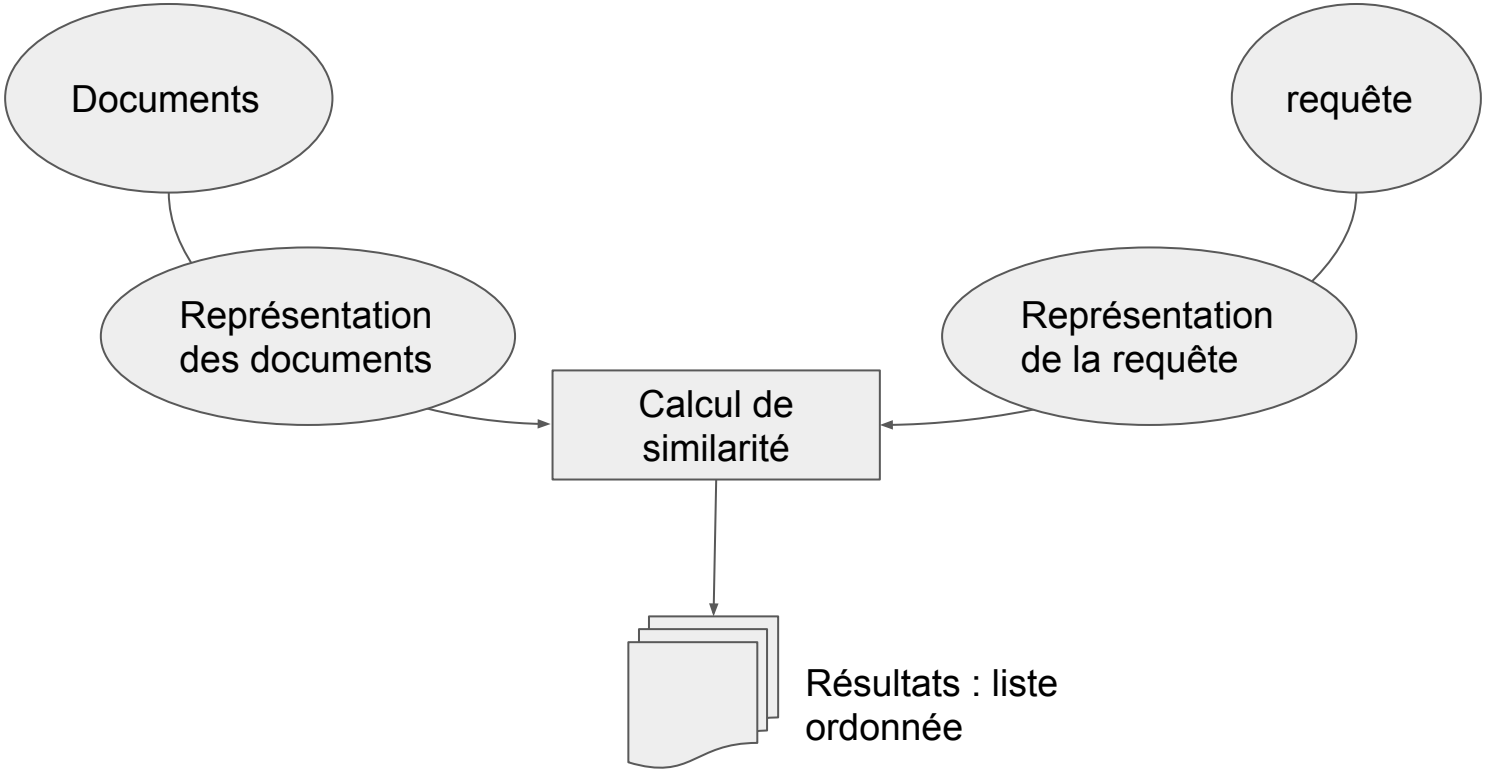
Recherche d'information



Recherche d'information



Recherche d'information



Représentation

Indexation plein-texte

- contenu du doc → représentation du doc
- choix des termes d'indexation : mots-formes

Tokénisation

- segmenter le texte en une liste de tokens (~ mots-formes)
  - token : suite de caractères entre 2 séparateurs (espace, ponctuation...)
  - pas si simple: c’est-à-dire qu’aujourd’hui, les pommes de terre des U.-S.-A. sont cultivées in vitro
- difficulté variable selon la langue (chinois, allemand...)

Représentation

Post-traitement

- exclusion des mots grammaticaux (prépositions, articles...)
  - utilisation de listes construites manuellement pour chaque langue
- option : racinisation
  - processus agressif basé sur des heuristiques
  - `compilers → compil, compilation → compil`

Similarité

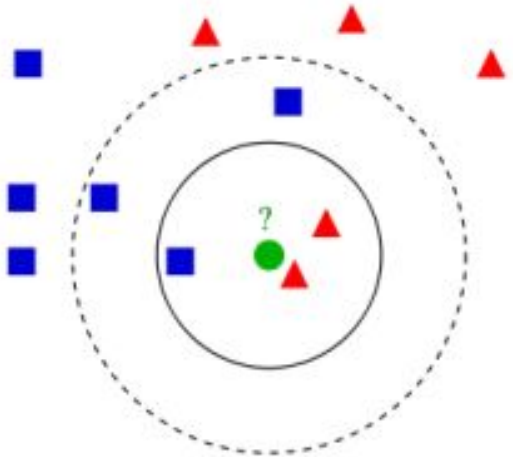
- pondération
  - TF-IDF : fréquence ds le doc vs. fréquence dans la collection
  - Okapi-BM25 : version moderne du TF-IDF [Robertson98]
- vecteurs (creux) [Salton71]
- similarité : cosinus, produit scalaire

Classification et RI

K-plus proches voisins

- apprentissage paresseux
- représentation RI
- similarité RI

⇒ pas d'optimisation de la similarité sur les données



Zoom : Datation de documents [DeFT 2012]

Tâche

- journaux 1800->1944
- OCR -> bruit

11 JUIN 1827.

On s'abonne rue des Prêtres Saint-Germain-l'Auxerrois, n° 17.

JOURNAL DES DÉBATS

POLITIQUES ET LITTÉRAIRES.

PORTUGAL.

Lisbonne, 26 *mai*.

(Correspondance particulière.)

Les courriers venant de Madrid, de Paris et de Vienne se sont succédé ces jours derniers. Le dernier nous apprend que le Prince Metternich n'a pas jugé à propos de se mettre en opposition ouverte avec l'Empereur D. Pedro, et qu'en conséquence l'infant D. Miguel va incessamment s'embarquer pour le Brésil.

La Princesse-Régente est partie avant-hier pour la maison de campagne qui lui avoit été préparée à Benfica; les deux princesses, ses sœurs, ne l'ont pas accompagnée, elles se sont au contraire rendues au palais Guelus, auprès de la Reine-Mère.

voter en sens contraire dans la prochain
de leurs places dans la Maison du Roi.

Lord Goderich a déclaré qu'il ne voul
de la couronne en répondant à cette
maître de conserver ou de renvoyer de s

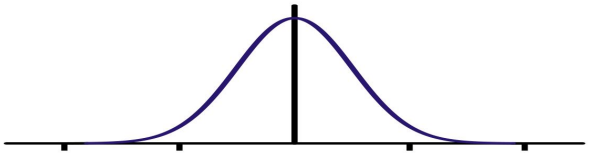
Lord Goderich a annoncé qu'afin qu
quelques jours fût de nouveau mis en d
le rapport du comité fût présenté mardi
poseroit la troisième lecture du bill de v

Dans la Chambre des Communes, le
au nom du Roi :

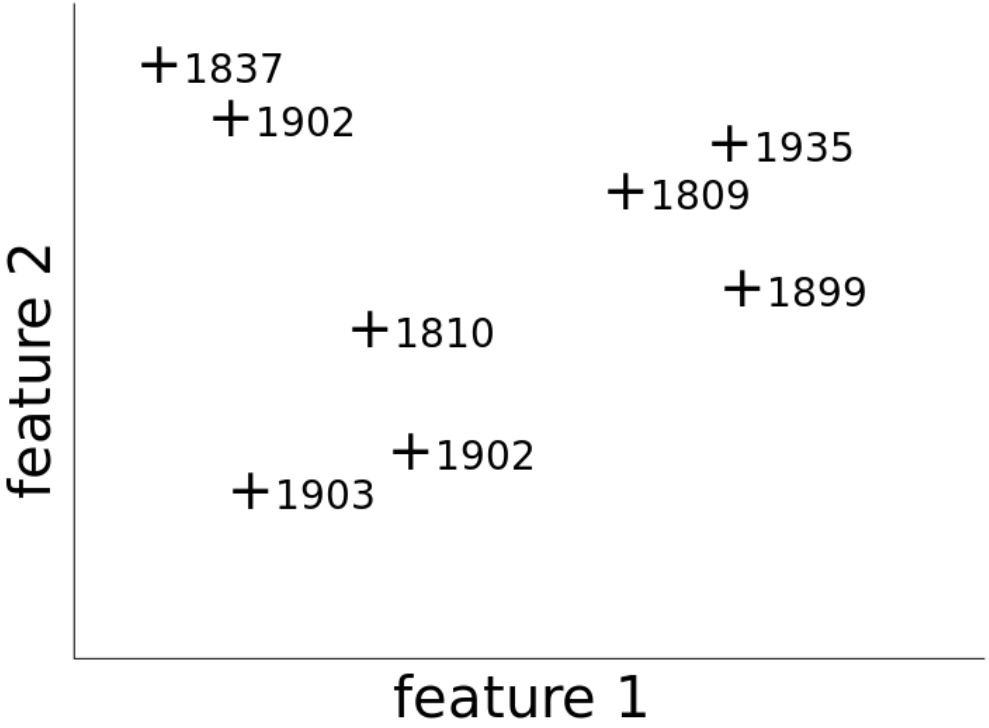
GEORGES ROI.

Kppv

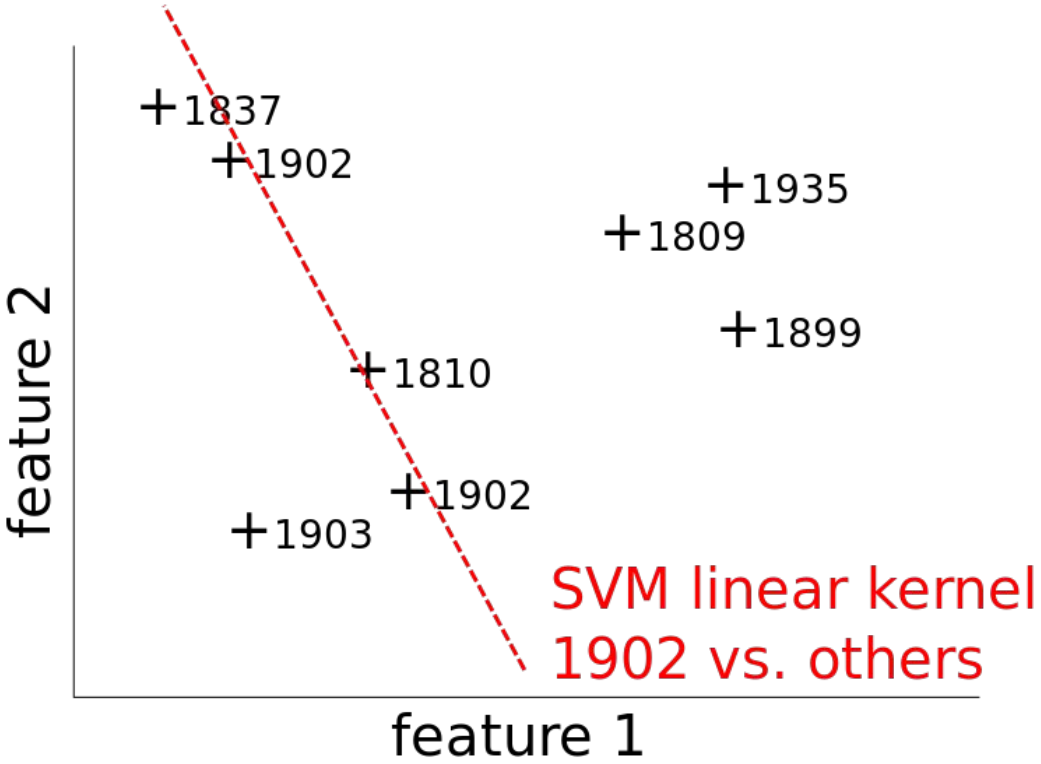
- Okapi-BM25
- vote sur les années adjacentes



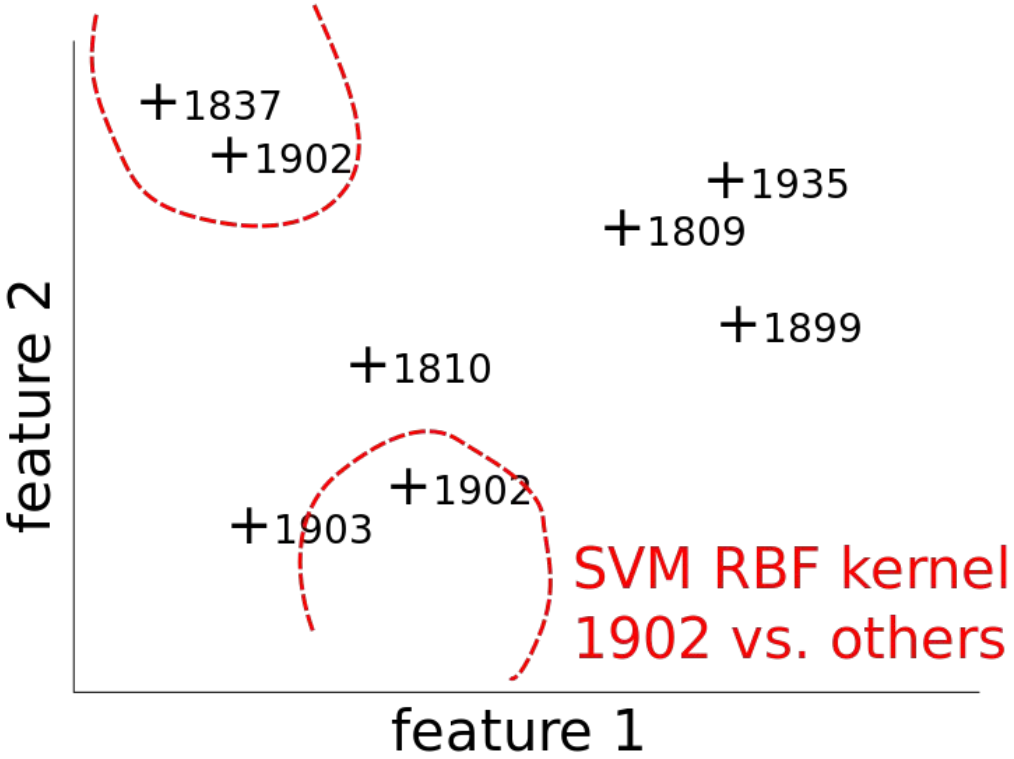
Intérêts du kppv



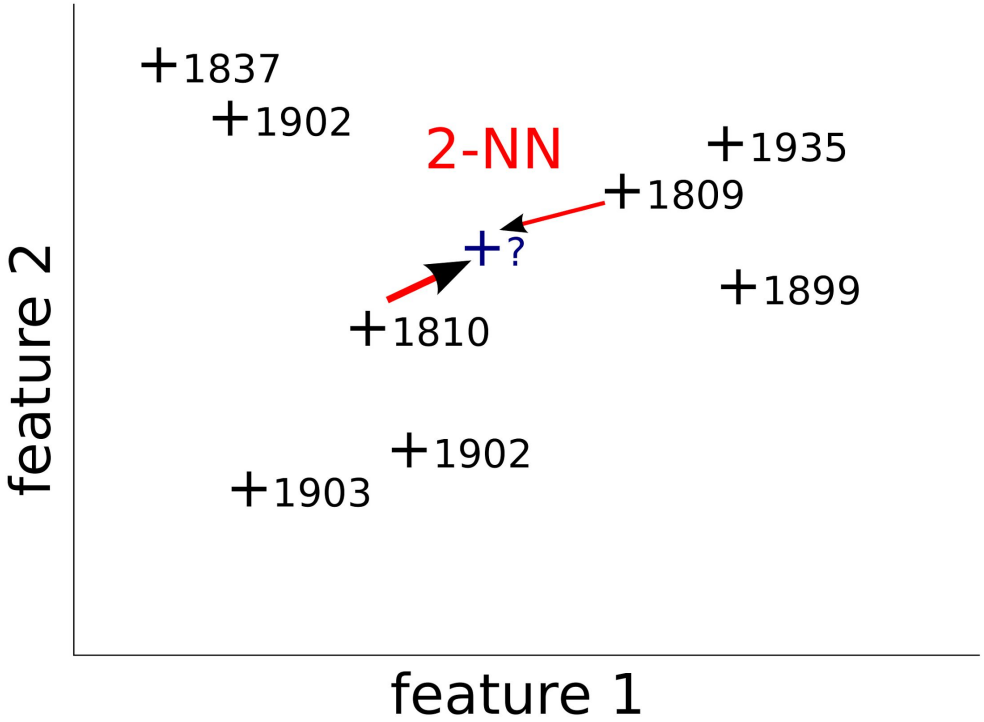
Intérêts du kppv



Intérêts du kppv



Intérêts du kppv



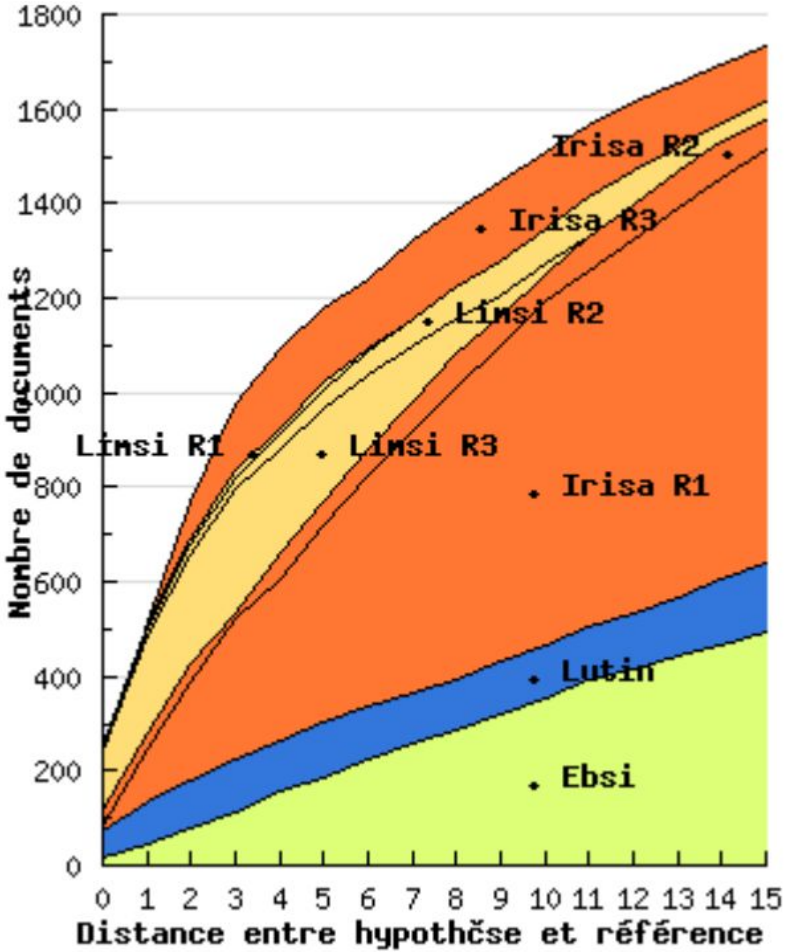
Résultats

Evaluation

- mesure tenant compte de la proximité avec la vérité terrain
- large gain par rapport aux autres méthodes

Autres avantages

- facilité à mettre en oeuvre
- facilité d'ajout d'exemples

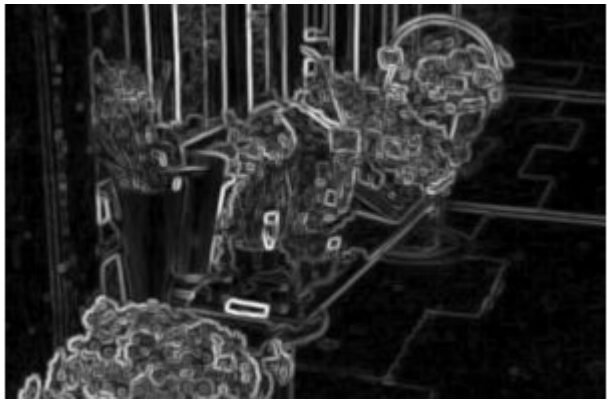
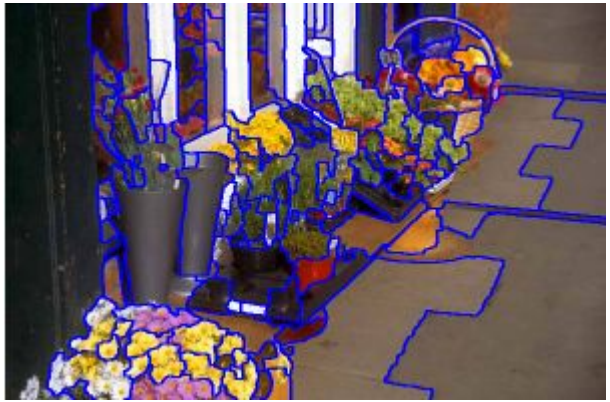


Zoom: Segmentation thématique

Segmentation d'émissions TV

- ~~indices visuels ou sonores~~
- reco de la parole -> flux de texte (bruité)
- détecter des portions cohérentes / changement de vocabulaire

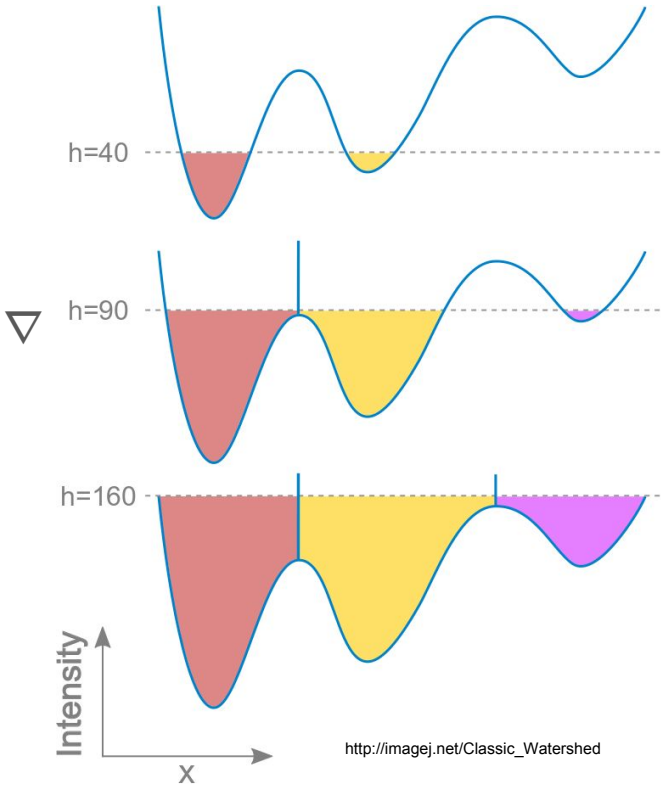
Analogie avec la segmentation d'image

- Segmentation des régions d'images, par leur contenu
- Simplement : les pixels voisins avec la même intensité sont groupés

⇒ en pratique: transformer l'image en calculant le gradient ∇

Analogie avec la segmentation d'image



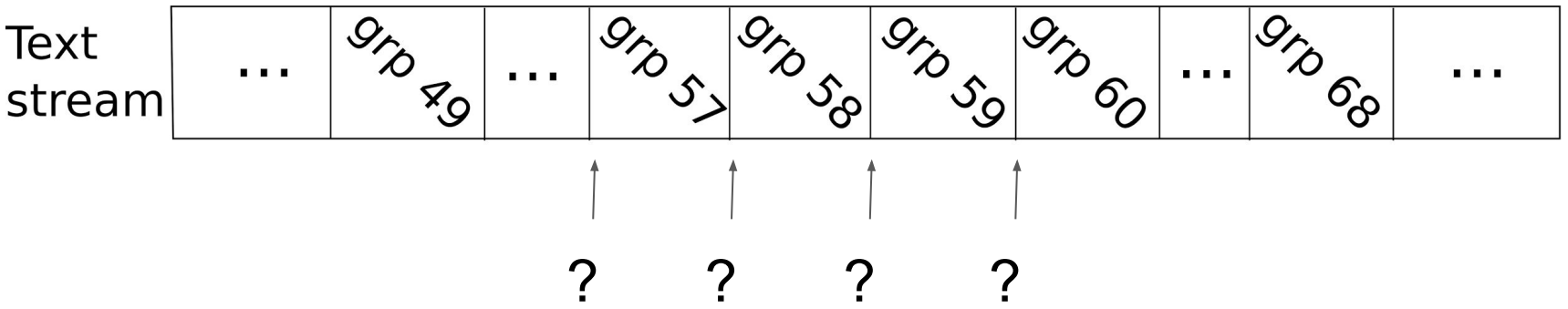
Watershed

- algo basique pour détecter des frontières dans l'image gradient
- verser de l'eau et construire des digues pour éviter le mélange d'eaux de différents bassins
- digues = frontières de régions

Retour à la segmentation de texte

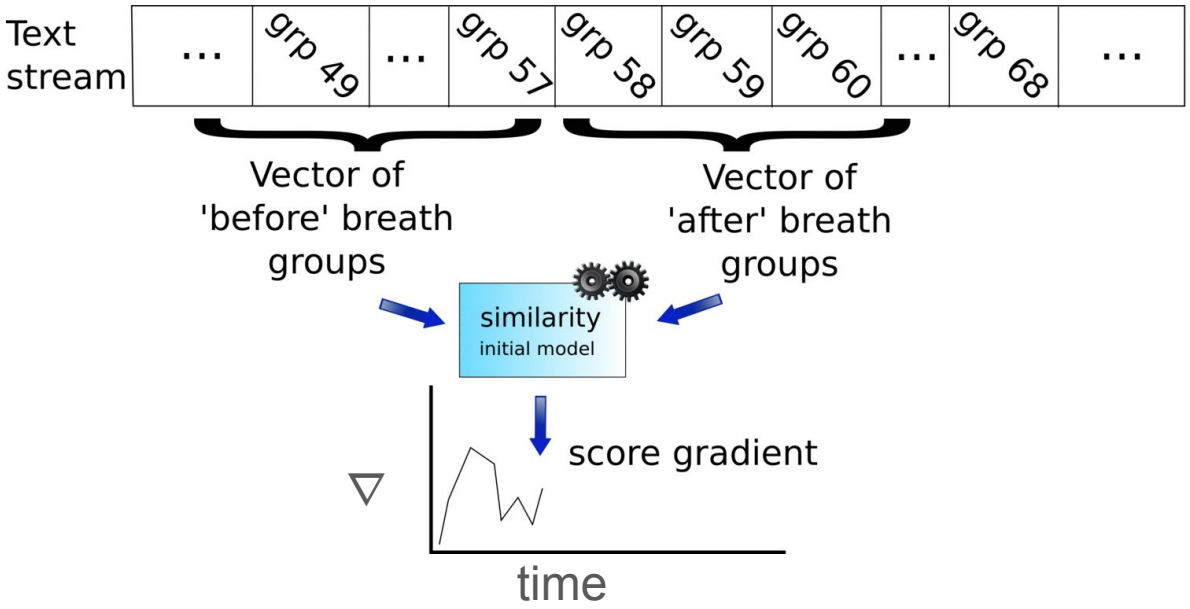
Sortie de reco de la parole

- groupe de souffle: segment de parole entre deux pauses ~ phrase
- *la chanteuse américaine* ~~*de dettes*~~ vedette *la plus réputée en france* ~~*dix mas combler*~~ *Liz McComb ce soir en concert à paris*



Gradient pour un flux de texte

- définir un gradient adapté au texte
- watershed pour trouver les limites des segments thématiques



Gradient pour un flux de texte

Représentation sac-de-mots

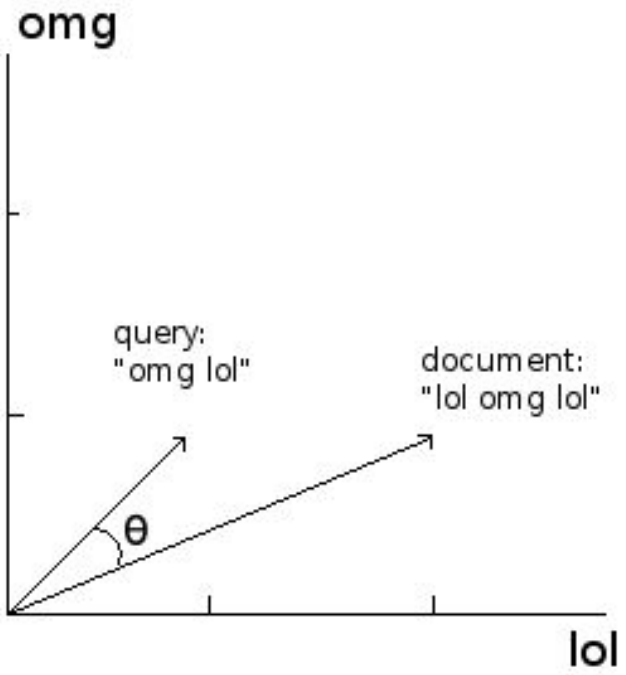
- groupe de souffle = ensemble de mots

Pondération

- TF: term frequent in the groups → more weight
- IDF: frequent in the whole stream → less weight
- more modern: Okapi [Robertson et al 98]

Modèle vectoriel [Salton 71]

- gr. de souffle = vecteur de poids
- similarité = cosinus/produit scalaire



Gradient pour un flux de texte

Before vs. After similarité

- TF-IDF/cosine ~ Text-Tiling [Hearst97]
- lexical chains, term repetition [Choi00, Reynar 00]
- language models [Utiyama et al. 01]

Problèmes

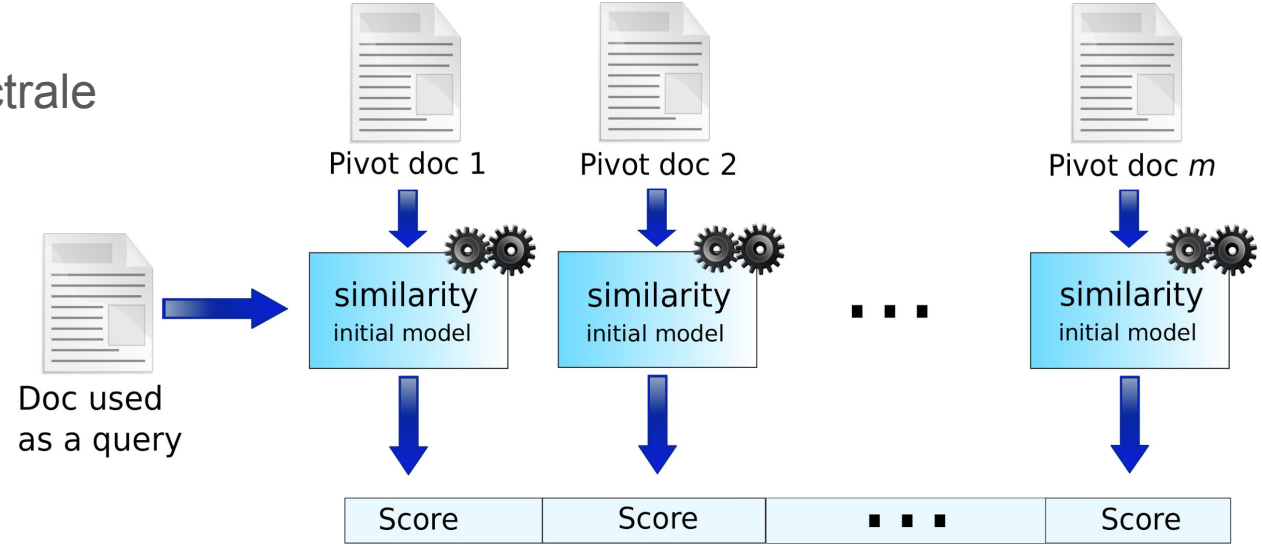
- Peu de mots répétés + erreur Speech2Text + petits segments

→ **similarité directe peu fiable**

Gradient pour un flux de texte

Vectorisation

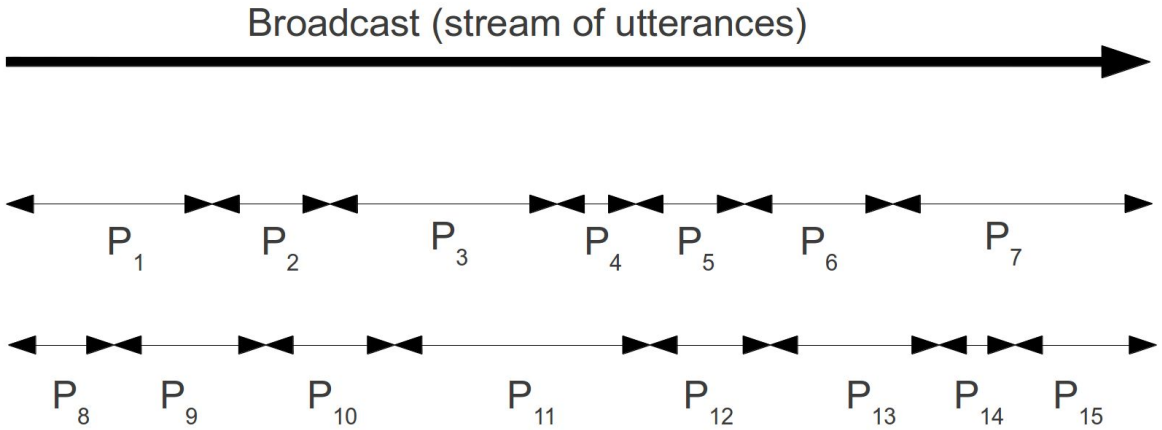
- représentation spectrale
- similarité 2nd ordre



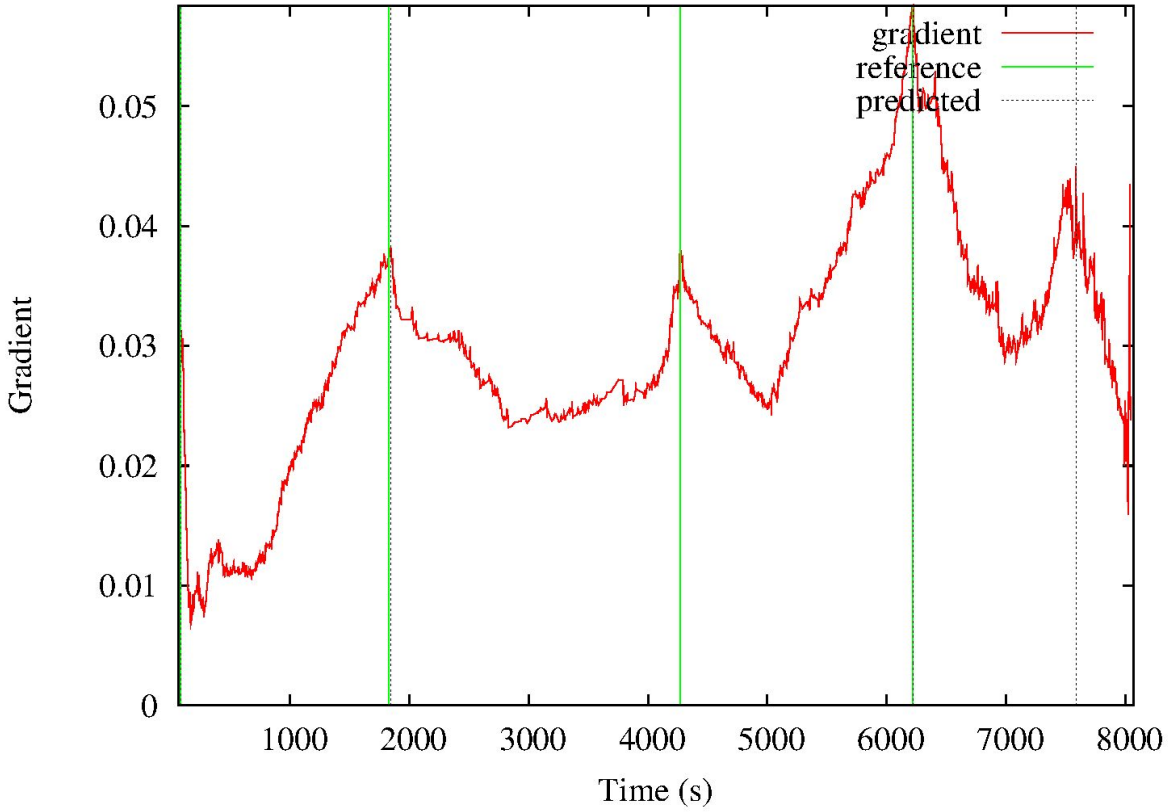
Gradient pour un flux de texte

Vectorisation

- permet de rapprocher des docs ne partageant pas les mêmes mots
- ici, choix aléatoire des pivots



Gradient pour un flux de texte



Évaluation

Données

- 2 collections d'emissions (fr): JT et reportages
- Differentes transcriptions: IRENE [Huet et al. 01], LIMSI [Gauvain et al. 02]

Cadre

- Comparaison avec vérité-terrain: Precision Rappel, F-mesure, WindowDiff
- Watershed + TF-IDF/Okapi/Vectorization
- Baselines:
  - historique: DOTPLOT [Reynar98], C99 [Choi00], TEXT TILING [Hearst97]
  - état-de-l'art: Utiyama [Utiyama01] (language modeling), Guinaudeau [Guinaudeau et al. 10] (Utiyama + semantic relations)

Évaluation

Reportages

- 12 *Envoyé spécial*, 16 *Sept-à-huit*
- peu de (grands) segments

Methods	P	R	F1	WD
Baseline	1.9	1.9	1.9	0.364
Utiyama and Isahara (2001)	75.3	73.6	74.4	-
DotPlot (Reynar, 2000)	49.49	49.49	49.49	0.2125
c99 (Choi, 2000)	57.42	57.42	57.42	0.1893
TextTiling (Hearst, 1997)	25.96	21.27	23.38	0.3456
TF-IDF + Watershed	59.32	60.93	60.12	0.1844
Okapi + Watershed	72.91	65.89	69.22	0.1428
Vectorization + Watershed	**77.98**	**72.57**	**75.18**	**0.1181**

Évaluation

Journaux TV

- 60 JT de *France 2*
- bcp de (petits) segments

Methods	P	R	F1	WD
Baseline	15.39	13.39	15.39	0.546
Utiyama and Isahara (2001)	57.6	61.4	59.44	-
DotPlot (Reynar, 2000)	36.42	36.42	36.42	0.4472
c99 (Choi, 2000)	50.25	50.25	50.25	0.3646
TextTiling (Hearst, 1997)	47.25	35.96	38.73	0.313
TF-IDF + Watershed	48.17	49.82	49.40	0.3421
Okapi + Watershed	64.06	56.49	60.04	0.2571
Vectorization + Watershed	**66**	**72.44**	**69.07**	**0.2269**

Évaluation

Impact des erreurs de transcription

- Word Error Rate: IRENE ~ 36%, LIMSI ~ 30%, manuel 0%

Methods	IRENE F1 / WD	Manual F1 / WD
Utiyama and Isahara (2001)	65.56/-	72.96/-
Guinaudeau et al. (2010)	68.94/-	73.30/-
TF-IDF + Watershed	60.73/0.2854	71.35/0.1978
Okapi + Watershed	63.38/0.2702	70.58/0.2075
Vectorization + Watershed	**69.44/0.2096**	**73.66/0.1851**

Évaluation

Leçons principales

- similarité RI → gradient pour flux de textes
- vectorisation utile pour comparer thématiquement des textes avec des vocabulaires différents

Perspectives

- segmentation hiérarchique
- semi-supervision: possibilité d'utiliser des marqueurs (~image)
- inclusion d'indices audiovisuels

Autres tâches (en compétition)

Extraction de terminologie [Quaero]

Appariement articles/résumés [DeFT11]

Attribution de mots-clés [DeFT12]

Analyse de sentiments [DeFT15, DeFT17]

Reconnaissance d'entités [BioNLP 13]

Indexation MeSH de documents médicaux [JRS12]

Détection de fake news [MediaEval16]

...

TAL pour la RI

Difficultés

Paraphrasage

- 2 énoncés différents ↔ 1 même sens
- rappel ↘

Ambiguïté

- 2 énoncés similaires ↔ 2 sens différents
- précision ↘

E. Voorhees :

It is not clear, [either] that NLP is required for some tasks that are closely related to ordinary retrieval.

TAL pour la RI

Phonétique
- Sons -> mots

Morphologie
- formation des mots

Syntaxe
- formation d'énoncés

Sémantique
- sens des mots / énoncés

Pragmatique
- contexte, connaissance du monde

TAL pour la RI

Phonétique
- Sons -> mots

Morphologie
- formation des mots

Syntaxe
- formation d'énoncés

Sémantique
- sens des mots / énoncés

Pragmatique
- contexte, connaissance du monde

Morphologie

Flexion

- `chien // chiens`
- paraphrasage

Outils

- racinisation, lemmatisation
- inclusion en RI : extension de requête ou conflation

Résultats

- racinisation efficace, même sur langues peu flexionnelles
- pas de gain à utiliser la lemmatisation [Moreau et al., 2007 ; Savoy, 2002]

Morphologie

Dérivation

- `compilation // décompilateur`
- paraphrasage

Outils

- racinisation, quelques bases morphologiques, quelques outils
- inclusion en RI : extension de requête ou conflation

Résultats

- gain variable selon les langues [Gaussier 99 ; Moreau et al., 2007]

Morphologie

Composition

- `maux d'estomac // stomachalgie`
- paraphrasage

Outils

- rares outils et bases en domaine spécialisé
- inclusion en RI : indexation sur les morphèmes

Résultats

- gains pour le domaine biomédical [Claveau et Kijak, 2013]

Zoom : Composition / domaine biomédical

paraphrasage : stomachalgie / gastrodynie / maux d'estomac
photochemotherapy ↔ photo / chemo / therapy

Intuition

- japonais comme langue pivot
- photochemotherapy ↔ 光化学療法
  - – photo ↔ 光 (’light’)
  - – chemo ↔ 化学 (’chemistry’)
  - – therapy ↔ 療法 (’therapy’)
- apprendre à aligner/traduire avec les kanjis

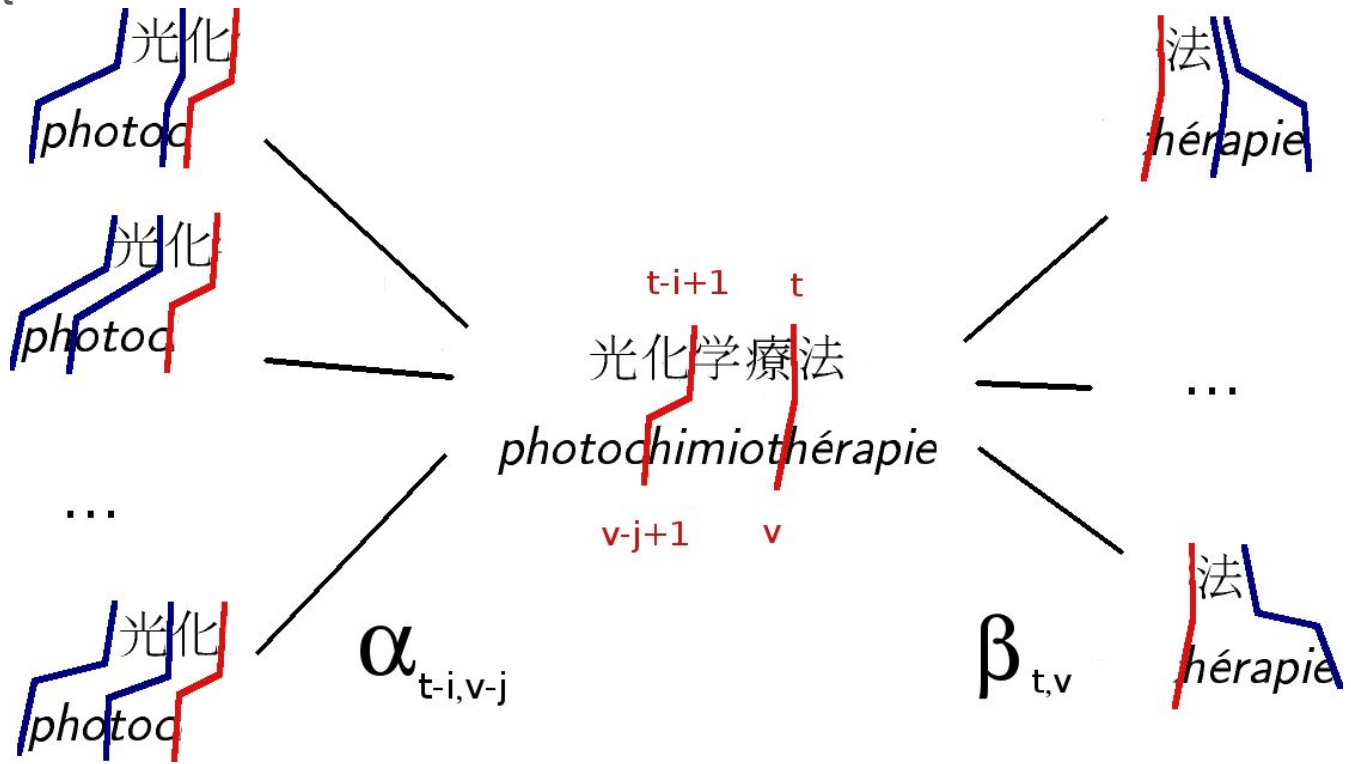
Composition / domaine biomédical

Alignement

- paires FR/JP tirées de l'UMLS
- algo forward-backward [Rabiner89, Jiampojamarn et al., 2007]
  - compte toutes les correspondances lettres/kanjis
  - Expectation: calcule une table de compte indiquant avec quel poids chaque alignement possible est rencontré, en s’appuyant sur la probabilité de cet alignement dans chaque paire considérée
  - Maximization: estime les probabilités d’alignement en s’appuyant à son tour sur la table de compte

Composition / domaine biomédical

Alignement



Composition / domaine biomédical

Termes alignés

卵子:ovo 形成:genesis;

高:hyper 炭酸症:capnia;

状:drepano 胞:cyt 性血:osis;

角膜全:kerato 移植:plasty;

下垂体:hypophys 切除:ectomy;

灼:caus 痛:algia;

...

Table de probabilités

上/ia; 0.00099

上嫌/euphor 4.95045e-05

上嫌/euphoria; 4.950495e-05

上炎/itis; 4.470132e-52

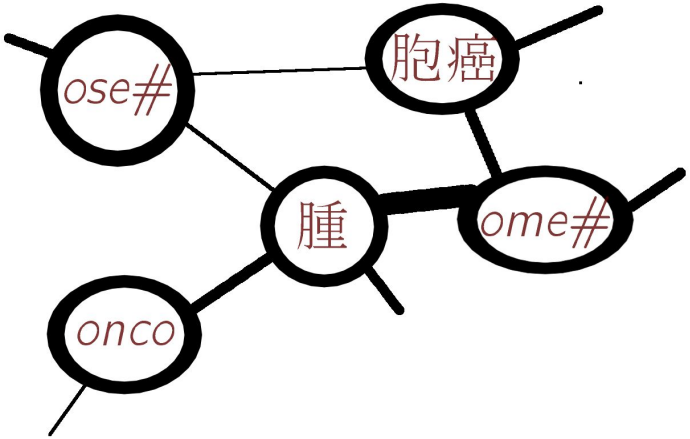
上狭窄/ostenosis; 5.59957e-23

上狭窄/stenosis; 7.716783e-17

上皮/carcino 2.568568e-311

...

Composition / domaine biomédical



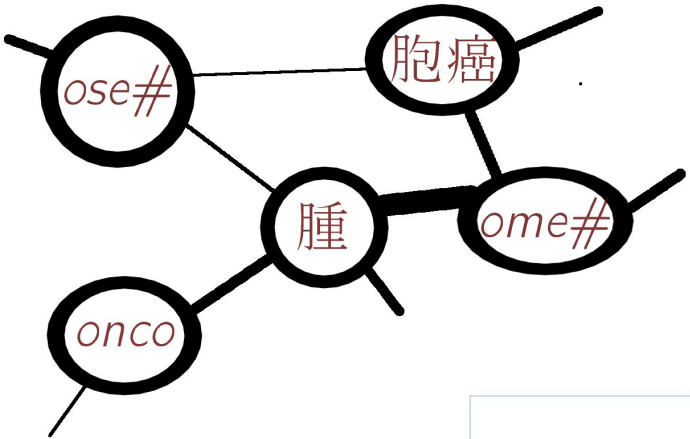
Représentation par kanjis du suffixe 'ome'

Composition / domaine biomédical

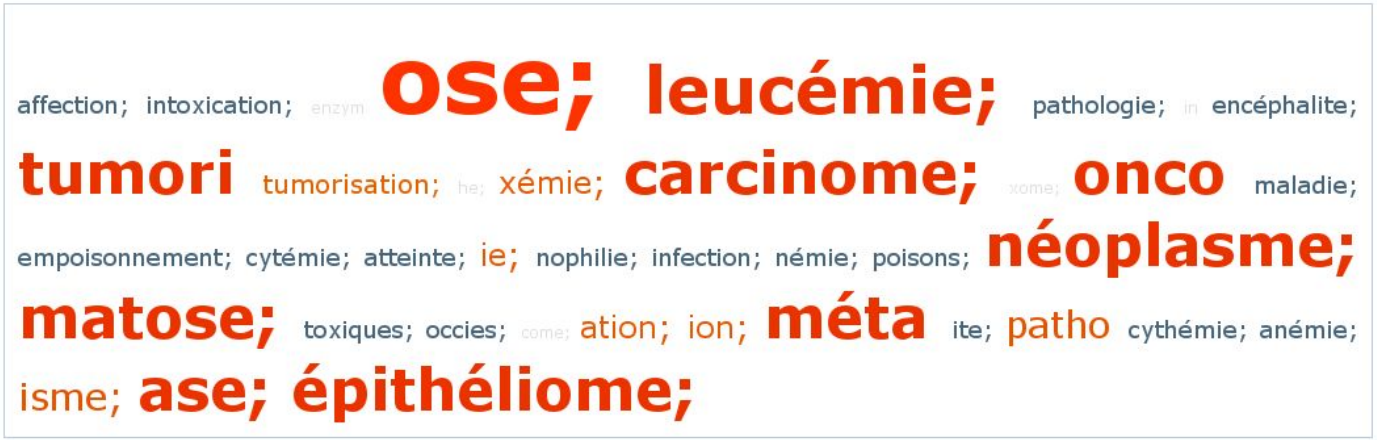
Sac de morphes Sac de kanjis

	baseline (BM-25 + stemming)	Morph-based System	Kanji-based System
MAP	29.93	33.94 (+13.4 %)	32.76 (+9.5 %)
IAP	31.74	35.55 (+12 %)	34.49 (+8.6 %)
R-prec	35.28	39.64 (+12.3 %)	38.59 (+9.4 %)
P@5	69.87	73.45 (+5.1 %)	71.70 (+2.6 %)
P@10	67.99	71.31 (+4.9 %)	*69.65 (+2.4 %)*
P@50	52.98	56.90 (+7.4 %)	55.24 (+4.3 %)
P@100	40.86	44.56 (+9.1 %)	43.39 (+6.2 %)
P@500	15.11	17.21 (+13.9 %)	16.92 (+12 %)
P@1000	8.72	10.10 (+15.86 %)	9.95 (+14.2 %)

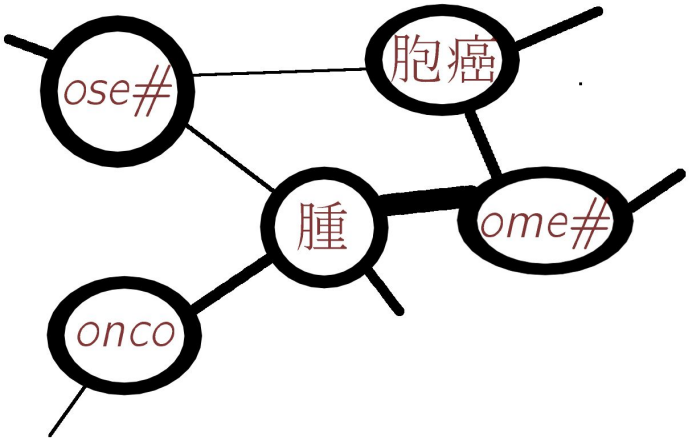
Composition / domaine biomédical



Voisins 1er ordre du noeud ome



Composition / domaine biomédical



Voisins 2nd ordre du noeud 'gatro'

Composition / domaine biomédical

Extension de requêtes

	baseline (BM-25 + stemming)	1st-order affinities	2nd order affinities
MAP	29.93	34.40 (+14.9 %)	28.74 (-3.9 %)
IAP	31.74	36.63 (+15.4 %)	30.80 (-2.9 %)
R-prec	35.28	39.92 (+13.2 %)	34.38 (-2.6 %)
P@5	69.87	71.76 (+2.7 %)	*68.65 (-1.7 %)*
P@10	67.99	70.46 (+3.6 %)	*66.20 (-2.6 %)*
P@50	52.98	56.30 (+6.7 %)	50.50 (-4.68 %)
P@100	40.86	44.69 (+9.4 %)	39.07 (-4.38 %)
P@500	15.11	17.98 (+18.9 %)	*15.01 (-0.64 %)*
P@1000	8.72	10.56 (+21.1 %)	*8.77 +0.66 %)*

Morphologie

Bilan

- les fruits bas sont déjà cueillis
- gains sur des contextes spécifiques (langues, domaines)

Mais aussi...

- agglutination (turc, allemand…) [Haddad et Bechikh Ali, 2014]
- voyellation (arabe) [Grefenstette et al., 2005]
- segmentation (chinois, japonais) [Peng et al., 2002]

TAL pour la RI

Phonétique
- Sons -> mots

Morphologie
- formation des mots

Syntaxe
- formation d'énoncés

Sémantique
- sens des mots / énoncés

Pragmatique
- contexte, connaissance du monde

Syntaxe

Partie du discours

- `il montre la porte // il porte une montre`
- ambiguité

Outils

- étiqueteur en parties-du-discours
- inclusion en RI : indexation forme_tag

Résultats

- pas de gain, voire dégradation
- liste de mots vides suffit ?

Syntaxe

Syntagmes

- phrasème, terme complexe, expression multi-mots
- `une pomme de terre tombée // une pomme tombée à terre`
- ambiguité

Outils et résultats

- outils statistiques, liste d'expressions, terminologies...
- inclusion en RI : indexation des expressions comme un tout
- quelques bons résultats [Acosta et al., 2011 ; Chapelle et Chang, 2011]
  - UMLS et MMTx [Aronson et Lang, 2010] pour le domaine biomédical [Shen et Nie, 2015]
  - logs de moteurs de recherche généralistes [Chapelle et Chang, 2011]

Syntaxe

Dépendances

- `information retrieval, retrieval of information, retrieve more information, information that is retrieved → retrieve+information`
- paraphrasage

Outils et résultats

- analyseur en dépendances
- inclusion en RI : très difficile, nécessite de nouveaux modèles
- quelques bons résultats [Maisonnasse et al., 2008 ; Gao et al., 2004]

TAL pour la RI

Phonétique
- Sons -> mots

Morphologie
- formation des mots

Syntaxe
- formation d'énoncés

Sémantique
- sens des mots / énoncés

Pragmatique
- contexte, connaissance du monde

Sémantique

L'effet Voorhees - match aller

- WordNet
- utilisation de la hiérarchie IS-A pour sélectionner le sens
- dégradation des résultats

E.M. Voorhees. 1993. *Using wordnet to disambiguate word sense for text retrieval.* In Proceedings of the 1993 ACM-SIGIR Conference on Research and Development in Information Retrieval.

Études postérieures

- Gains légers [Zhong et Ng, 2012]

Sémantique

Désambiguïsation

- `avocat // avocat`
- ambiguïté

Outils

- *Word Sense Disambiguation* : lexiques et outils
- inclusion en RI : indexation des formes_sens

Résultats

- premières études négatives → effet Voorhees
- coût > gain attendus

Sémantique

Relations sémantiques

- `bicyclette // vélo`
- paraphrasage

Ressources

- lexiques/théausurus/terminologies/dictionnaires existants
- modèle de type *Latent Semantic Indexing*
- sémantique distributionnelle
- inclusion en RI : extension de requête ou modèles plus complexes

Résultats

- premières études négatives → effet Voorhees

Sémantique

L'effet Voorhees - match retour

- WordNet
- sélection manuelle (!) des synonymes
- résultats nuls : aucune amélioration

E.M. Voorhees. 1994. *Query expansion using lexical-semantic relations*. In Proceedings of the 17th ACM-SIGIR Conference.

Études postérieures

- Voir 2e partie

Sémantique

Bilan

- résultats variés selon les études
- mauvaise réputation due à des expériences dépassées
- inclusion relativement facile dans les systèmes des RI

Mais aussi…

- traduction dans la même langue [Gao et al 2010]
- nouvelles approches neuronales, à surveiller [Piwowarski et al. 2015]

Zoom : RI et lexique distributionnel

Qu'est-ce ?

- à chaque mot d'entrée sont associés des mots sémantiquement proches
- repose sur l'hypothèse distributionnelle :

  You shall know a word by the company it keeps [Firth, 1957]
- similarité entre contextes des mots au sein de gros corpus

Construction

- sujet ancien [Grefenstette 94]...
- …toujours actif grâce à la disponibilité de données [Lin 98, Curran & Moens 02]

Lexique distributionnel

En pratique

- nombreuses façons de calculer les similarités [Broda et al. 09, Yamamoto & Asakura 10]
- notion de proximité sémantique variée [Budanitsky and Hirst 2006; Adam et al., 2013]
  - synonymie
  - (autres) relations paradigmatiques
  - relations syntagmatiques

Lexique distributionnel

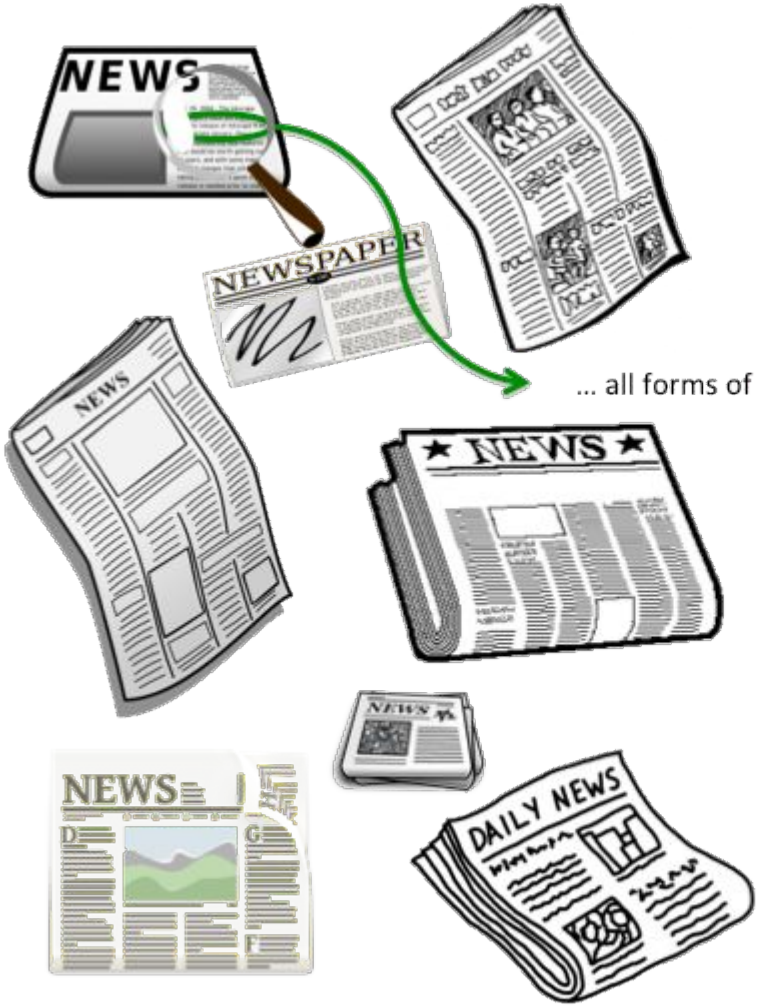
Évaluation intrinsèque

- comparaison avec des lexiques de référence :
  WordSim 353, WordNet, Moby, TOEFL, BLESS [Baroni & Lenci 2011]
- évaluation directe, mais questionne la référence

Évaluation extrinsèque

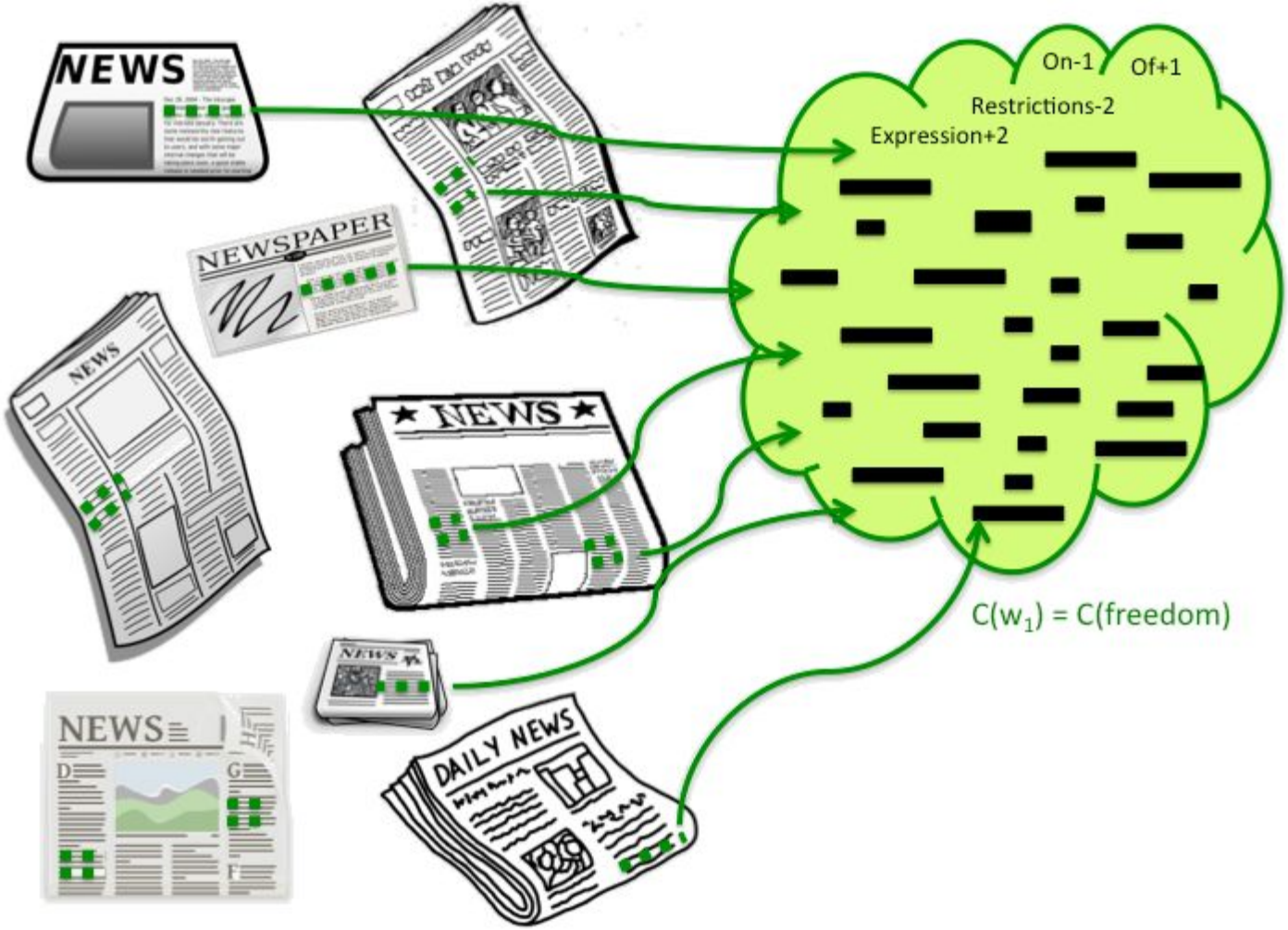
- évaluation indirecte via une application tierce
  - substitution lexicale [McCarthy & Navigli 2009]
  - tâche de recherche d'information [Besançon 99]
- pas de distinction entre création du thesaurus et la tâche

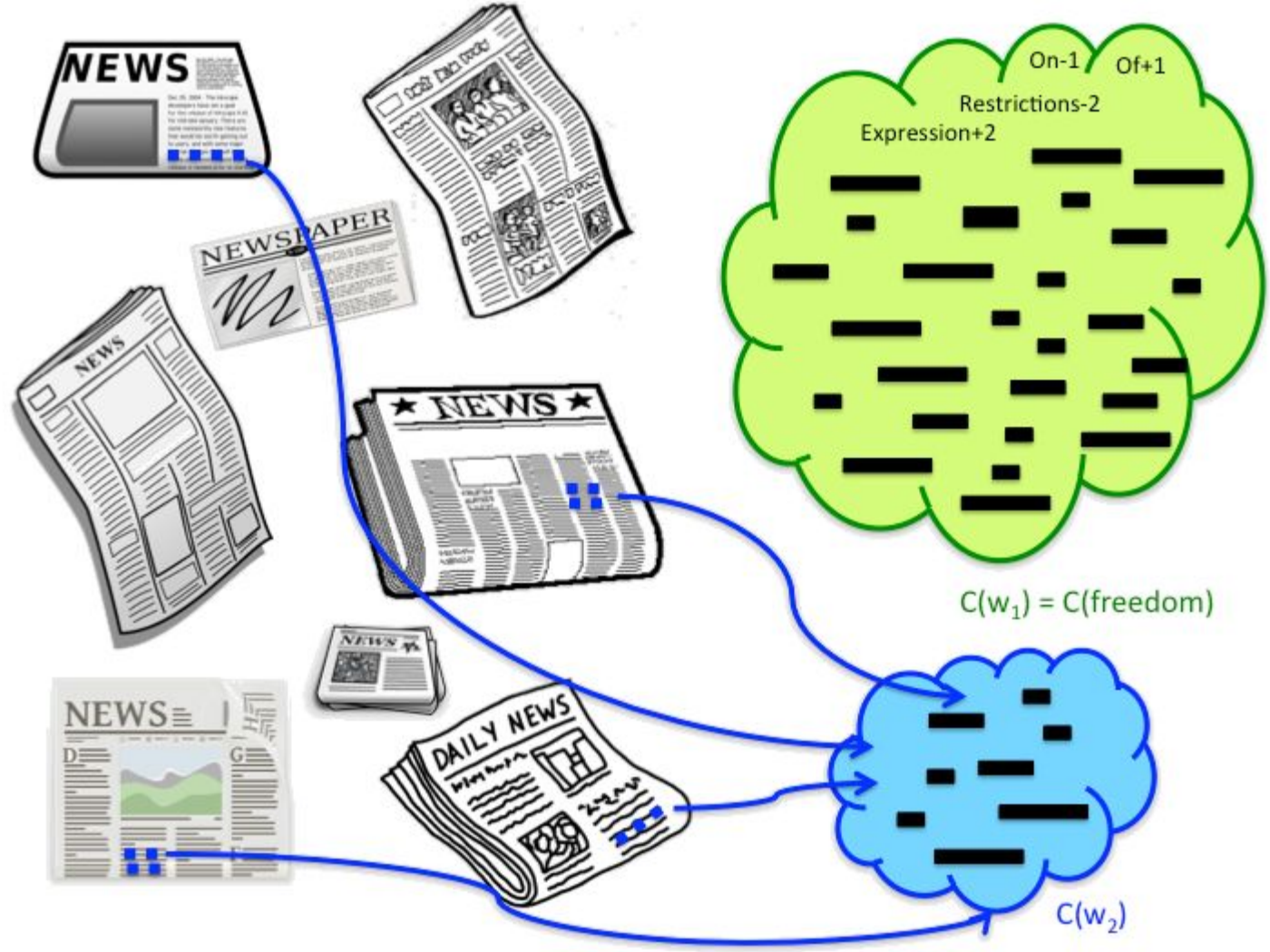

NEWS
NEWSPAPER
NEWS
NEWS
NEWS
NEWS
DAILY NEWS



… all forms of restrictions on **freedom** of expression, threats…

w_1-2 w_1-1 **w_1** w_1+1 w_1+2


NEWS
NEWSPAPER
NEWS
NEWS
NEWS
DAILY NEWS
On-1
Of+1
Restrictions-2
Expression+2
C(w1) = C(freedom)


NEWS
NEWSPAPER
NEWS
NEWS
NEWS
DAILY NEWS
On-1
Of+1
Restrictions-2
Expression+2
C(w1) = C(freedom)
C(w2)

Principe central [Claveau, Kijak, Ferret 2014]

Similarité avec des modèles RI

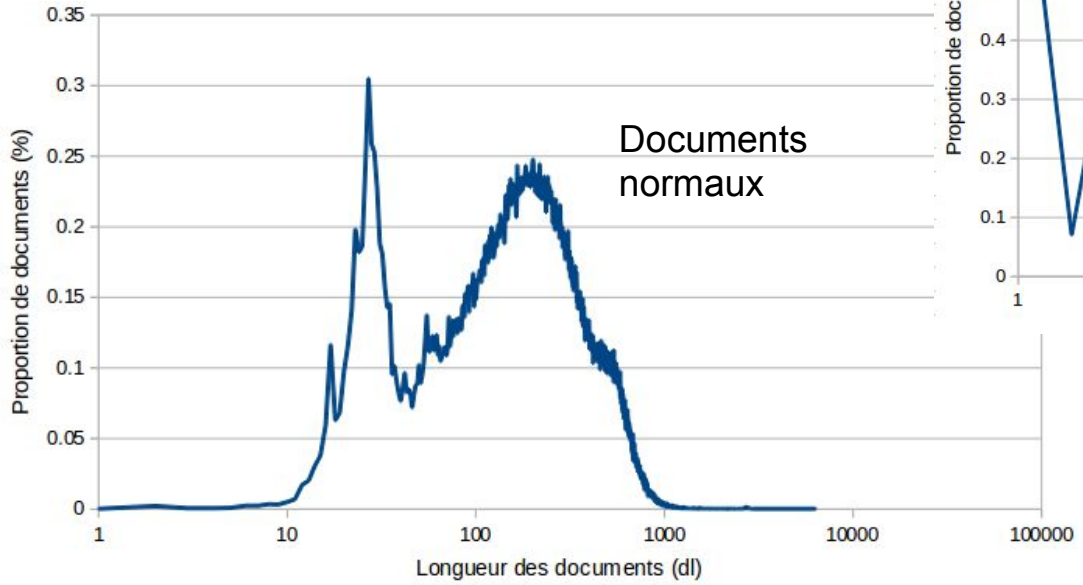
- contextes d'un mot → document
- contextes d'entrée → requête
- proximité distributionnelle → similarité de type RI

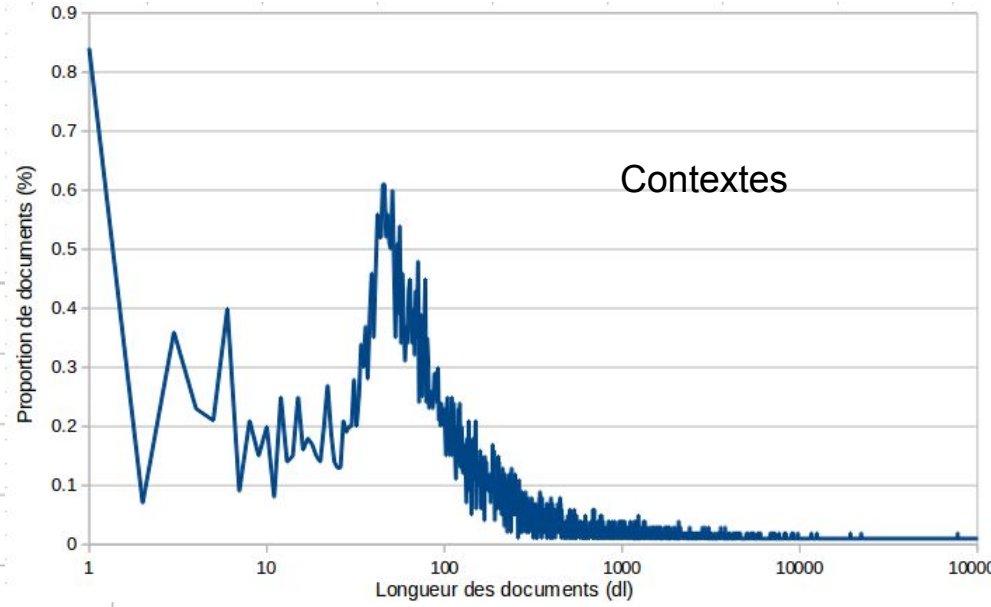
Modèles

- Hellinger, TF-IDF/cosinus, Okapi [Robertson et al. 98] adjusted with more weight to IDF
- LSI, RP, LDA avec plusieurs nombres de dimensions
- LM avec lissage Dirichlet ou Hiemstra (différents μ et λ)

Limite de l'analogie

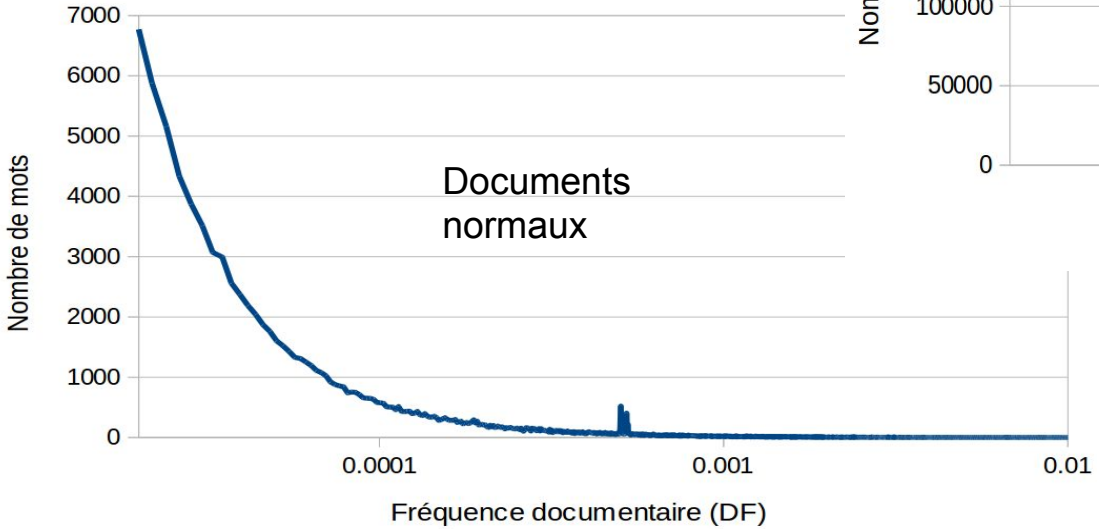
Longueur des documents

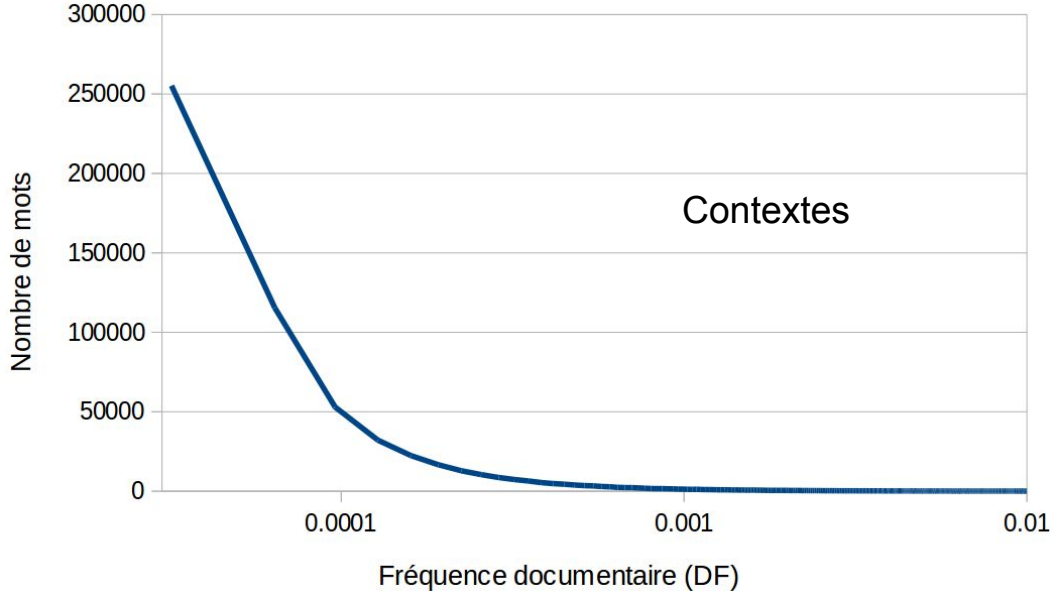




Limite de l'analogie

Fréquence documentaire





Évaluation par ressources externes

Ressources d'évaluation

WordNet [Miller 90]

- (quasi-)synonymes
- 3 mots liés en moyenne pour 10 000 noms

Moby [Ward 96]

- relations plus complexes
  - co-hyponymie / hyponymie : abolition-annulment, cataclysm-debacle
  - co-hyponymie / hyperonymie: abyss-rift, algorithm-routine
- 50 mots liés en moyenne pour 9000 noms

WN + Moby

- W+M: 38 mots liés en moyenne pour 14 000 noms

Protocole d'évaluation

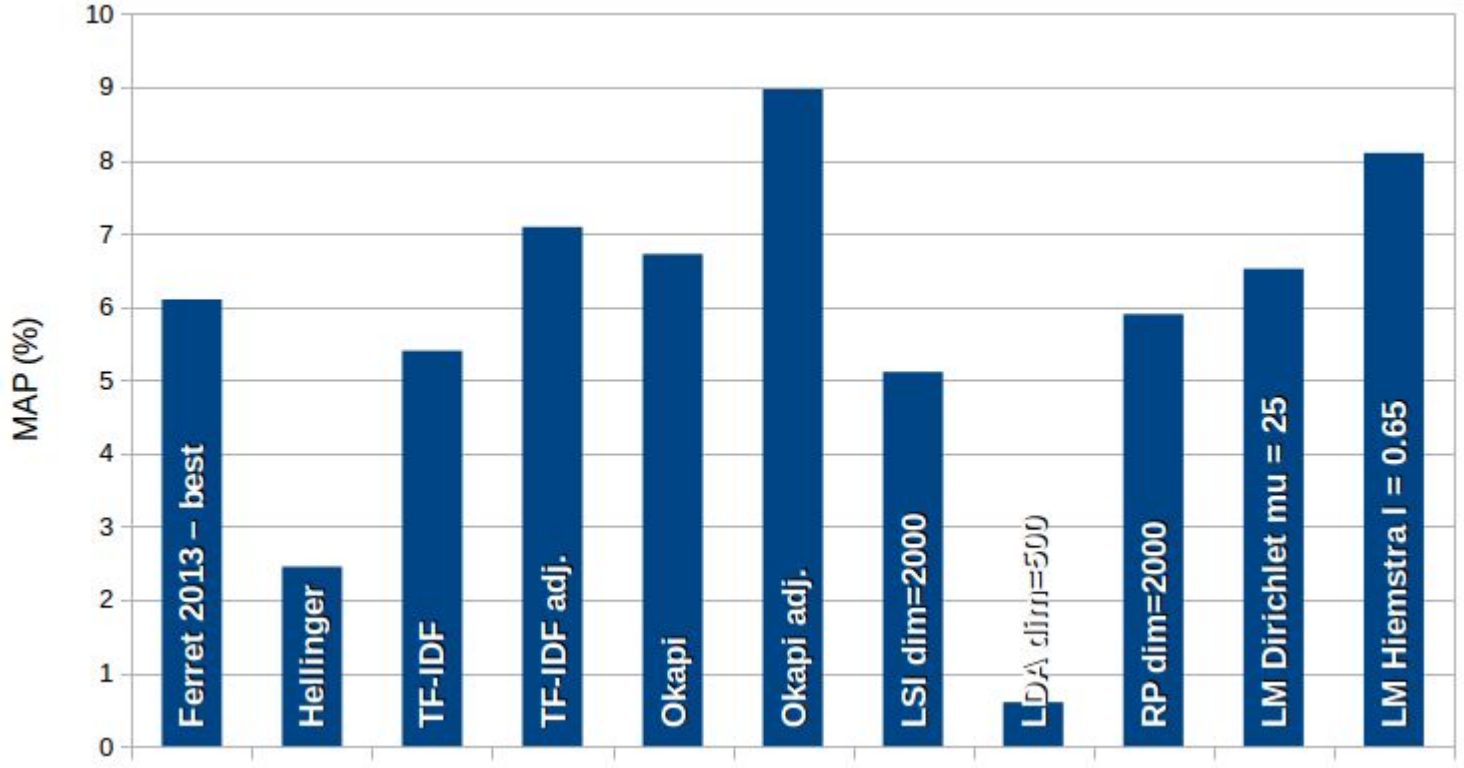
Corpus

- AQUAINT2: articles de presse, ~ 380 millions mots
- 25 000 noms uniques de freq ≥ 10

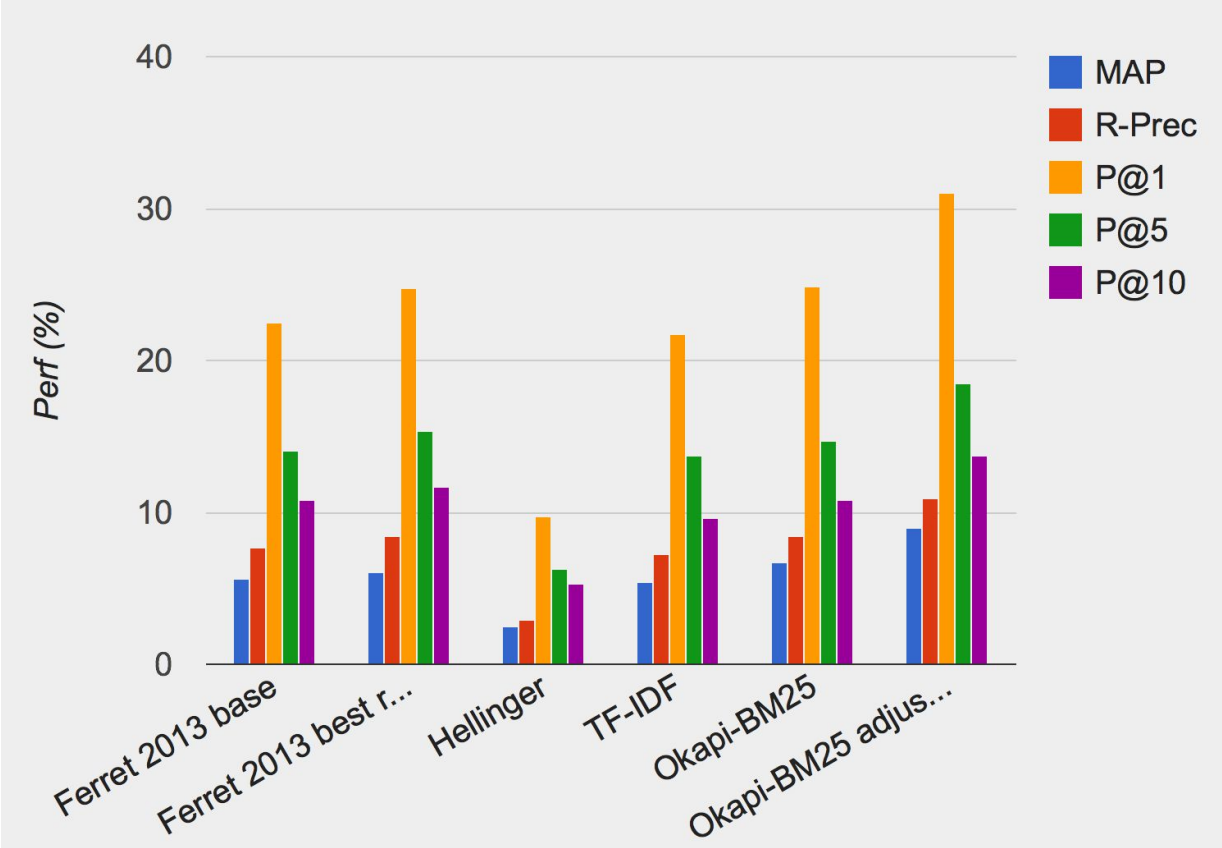
Expérience

- pour tous les noms du corpus : trouver les voisins (similarité de contextes)
- comparaison : lexique distributionnel construit vs. lexiques de référence
- MAP, R-Prec, précision sur les 5, 10, 20... premiers voisins
- système état de l'art : Ferret 2013

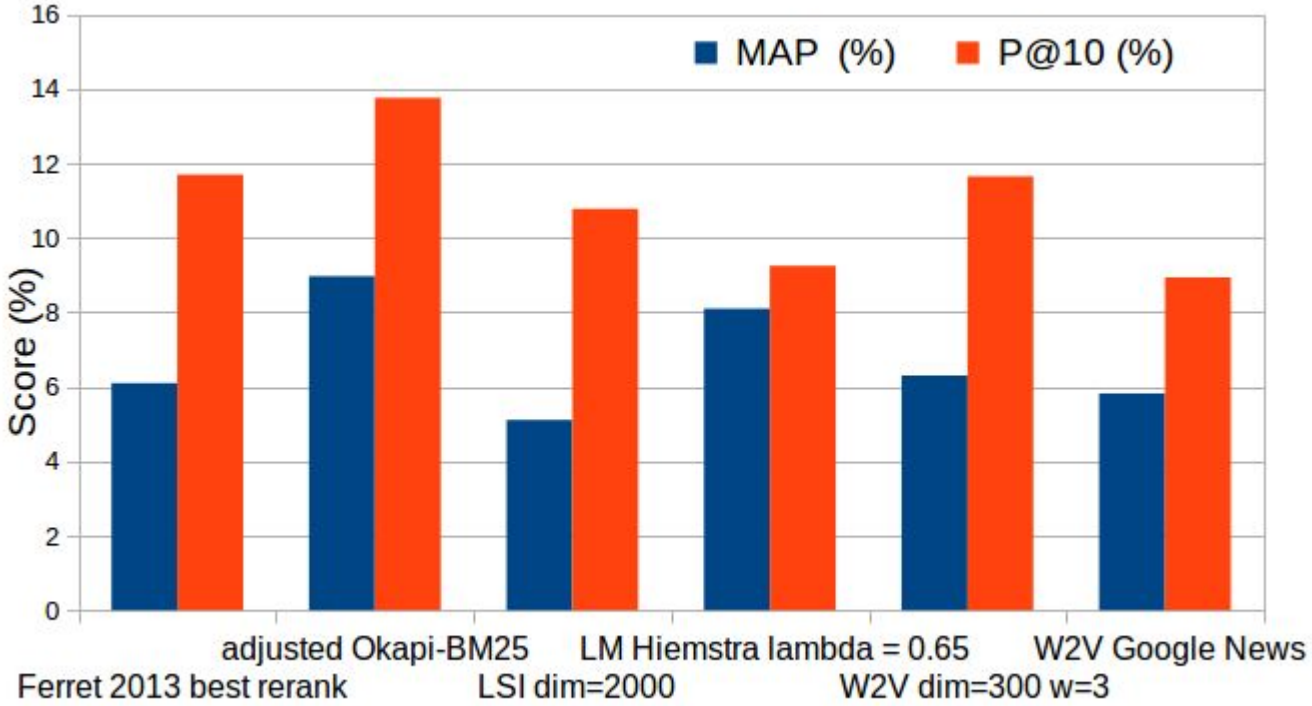
Performances globales



Performances globales



Performances globales



Performances globales

Premiers commentaires

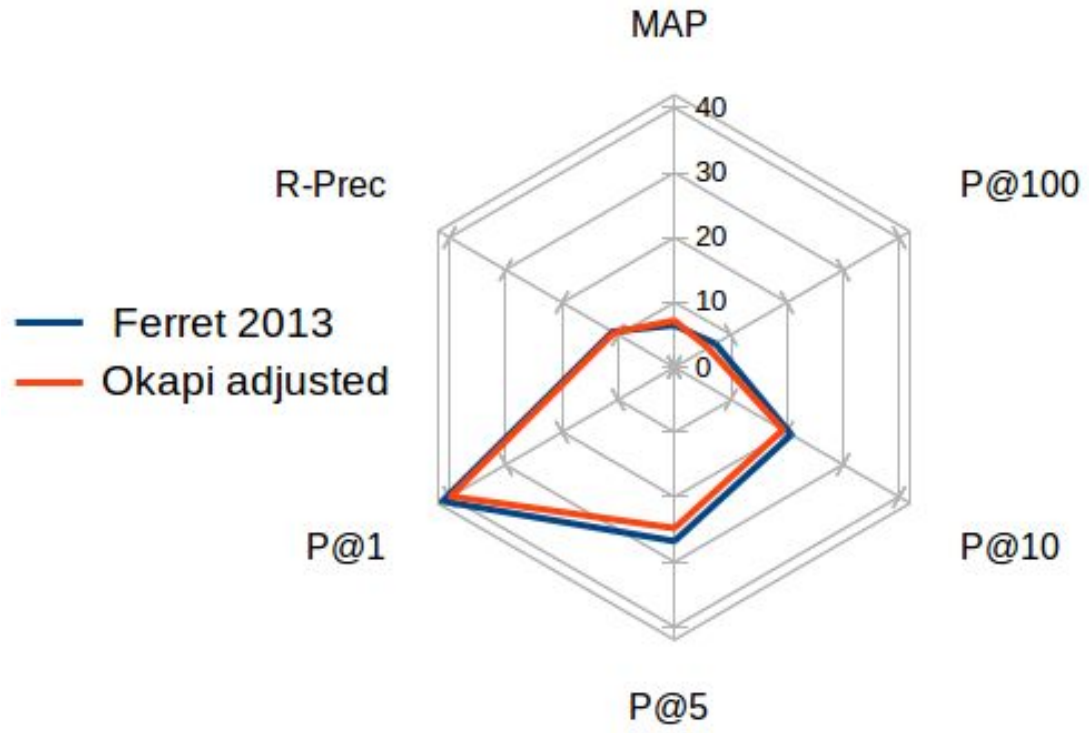
- MAP basse pour toutes les approches → tâche difficile ?
- approche RI fonctionne bien par rapport à l'état-de-l'art

Modèle par modèle

- Okapi et LM Hiemstra donnent les meilleurs résultats
- les modèles/paramètres montrent l'importance des mots de contextes discriminants

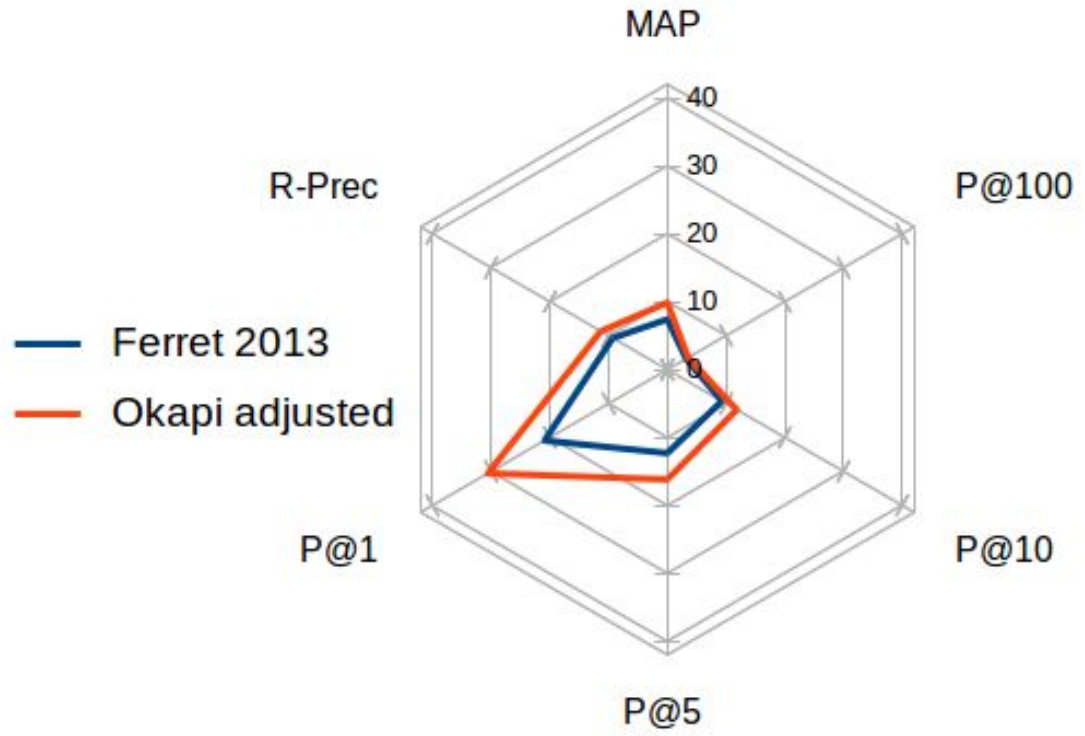
Performances selon la fréquence de l'entrée

Nb d'occurrences > 1000



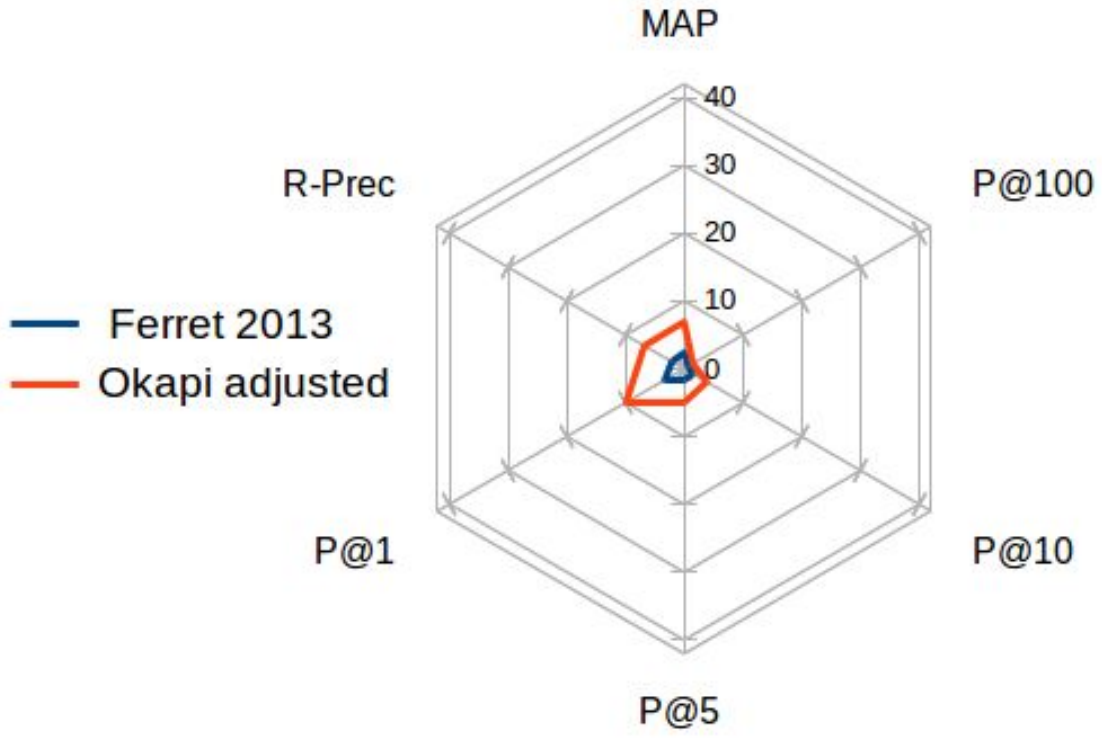
Performances selon la fréquence de l'entrée

100 < Nb d'occurrences < 1000



Performances selon la fréquence de l'entrée

Nb d'occurrences < 100



Autre évaluation: SimLex999 [Hill et al. 2014]

Données

- degré d'association plutôt que décision brutale
- score associé à des paires de mots (N, V, Adj) selon leur "force d'association"
  - Scores donnés par un groupe d'annotateurs avec peu d'instruction sur ce qu'est une "association"
  - Seulement des termes communs/fréquents → biais possible

Évaluation

- corrélation de Spearman entre liste ordonnée par les humains et la machine

Autre évaluation: SimLex999

	Adj Okapi	W2v dim=300 w=3	W2v GoogleNews
Spearman corr.	0.4511	0.3691	0.4419

- loin d'être parfait mais meilleur qu'attendu
- repr. spectrale > w2v sur les mêmes données d'entraînement

Autre évaluation: SemEval [McCarthy & Navigli, 2009]

Substitution lexicale

- trouver des mots similaires dans un contexte spécifique

Mise en oeuvre

- 10 plus proches voisins dans le thésaurus, qq soit le contexte
- évaluation Précision P@10

Méthodes	P (%)
Adj. Okapi	22
w2v dim=300 w=3	19
w2v GoogleNews	23.5

Bouclons la boucle: RI → TAL → RI

Évaluation par extension de requête

- utiliser les lexiques sémantiques pour améliorer un système de RI
- ajouter à chaque nom de la requête ses n plus proches voisins

Contexte expérimental

- collection Tipster : 170 000 articles de presse, ~ Aquaint
- système de RI : Indri (ML+réseau d'inférence, paramétrage standard)
- utilisation du thésaurus distributionnel pour étendre les requêtes
- baseline : idem avec les lexiques de référence

Évaluation par la RI

Exemple d'extension de requête

Requête :

```
coping with overcrowded prisons
```

Requête pour Indri :

```
#combine( coping with overcrowded #syn( prisons prison ) )
```

Requête étendue :

```
#combine( coping with overcrowded #syn( prisons prison inmate
inmates jail jails detention detentions prisoner prisoners detainee
detainees ) )
```

Évaluation par la RI

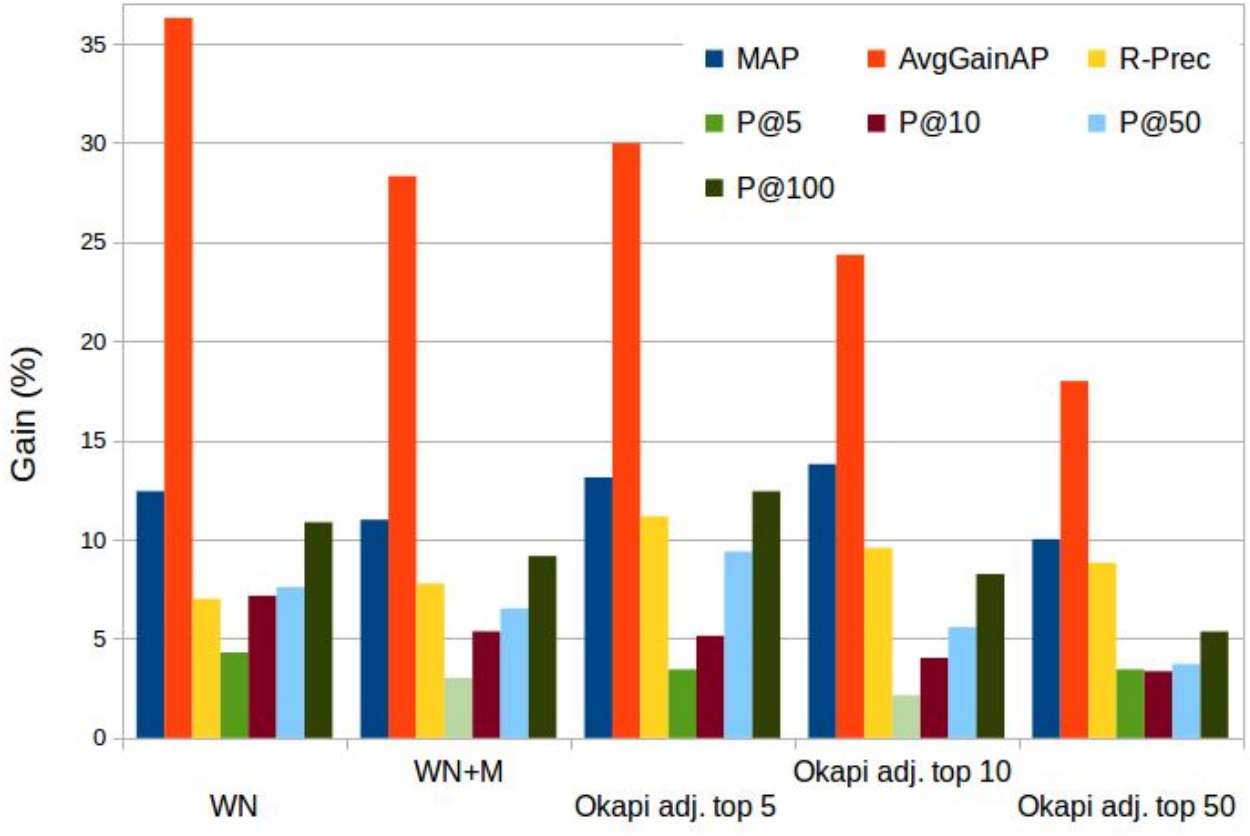
Messure de performances

- comparaison des résultats avec et sans extension
- mesures de RI standard + AvgGainAP

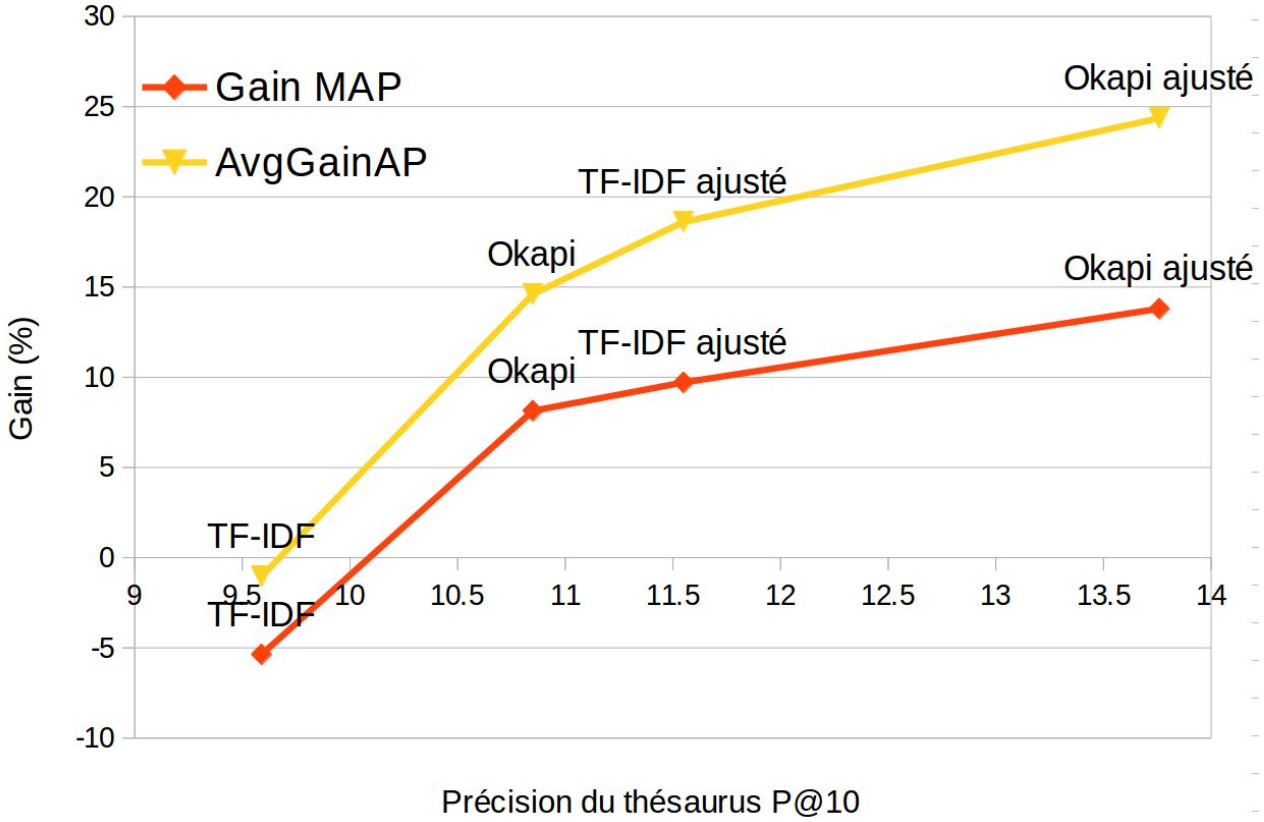
AvgGainAP vs. gain en MAP

- gain en MAP = gain sur les AP moyennée sur les requêtes
- les gains en MAP cachent des variations importantes sur certaines requêtes
- AvgGain = moyenne des gains des obtenus à chaque requête
  - AvgGainAP élevé → AP de la majorité des requêtes est améliorée
  - cas typique : gain en MAP mais faible AvgGainAP → gain important en AP sur qq requêtes

Résultats de l'extension de requête



Relation avec la qualité du thésaurus ?



Ressource vs. tâche

Précision intrinséque / extrinséque

Pas pire...

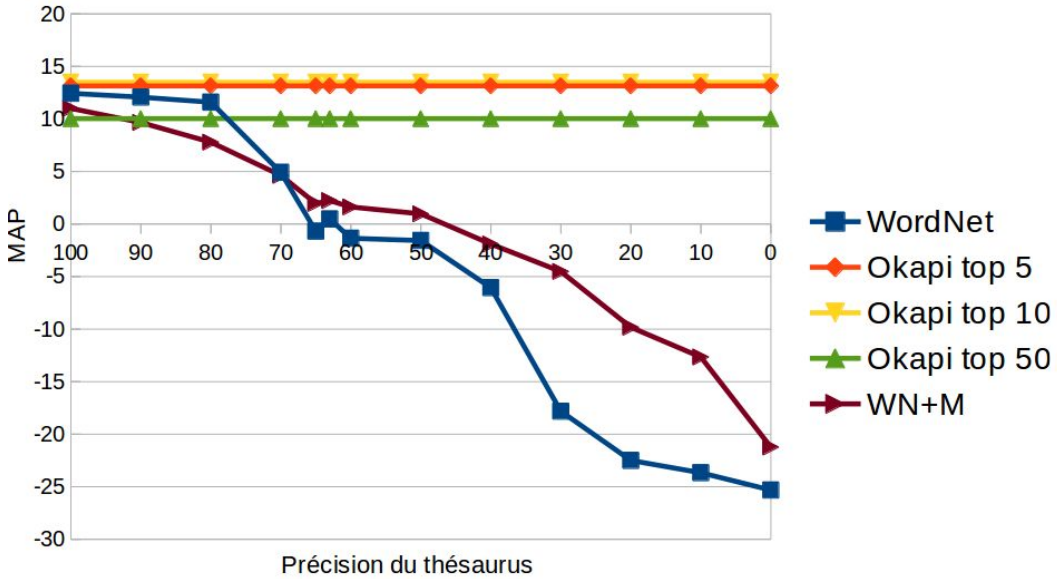
- performances comparables entre lexiques distributionnels et WN/Moby
- précision intrinsèque correlée mais très sévère

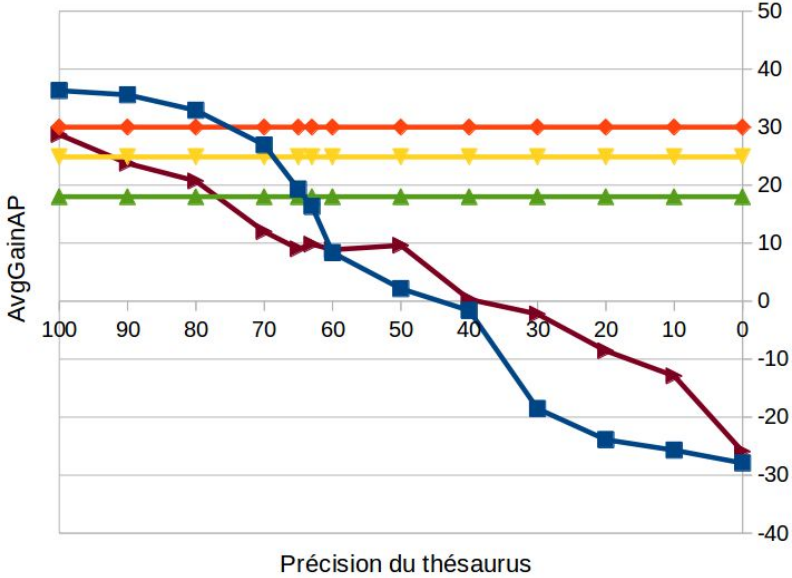
Generation de lexiques avec une précision contrôlée

- dans WN/Moby, remplacer 10%, 20% ... 100% des voisins par des mots au hasard
- refaire l'extension de requêtes avec ces références bruitées

Résultats

Précision contrôlée artificiellement





Gain en MAP (gauche) et AvgGainAP (droite) pour des lexiques avec une précision contrôlée artificiellement

Ressources incomplètes ?

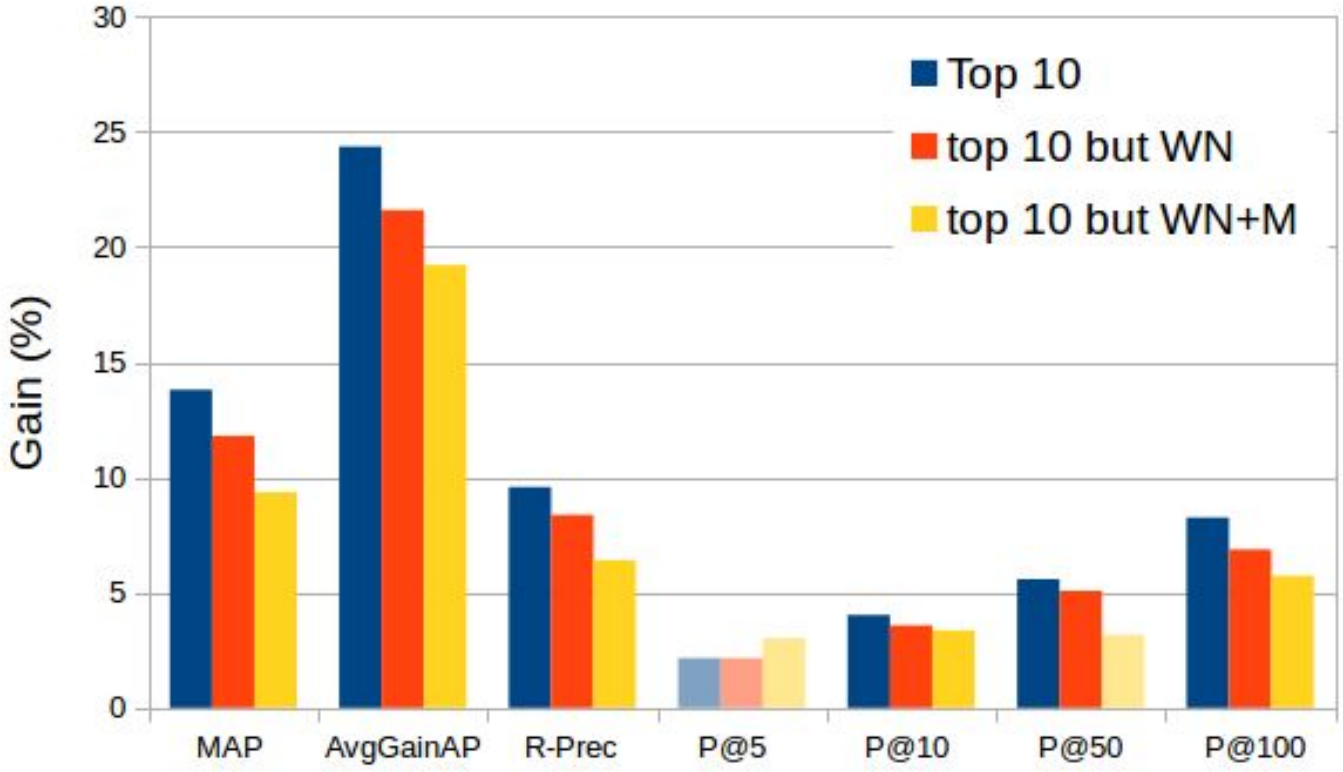
Faux voisins sémantiques ?

- mots absents des références ne sont pas toujours des erreurs
- à partir des 10 voisins de `prison`, ceux absents de WN+Moby sont:

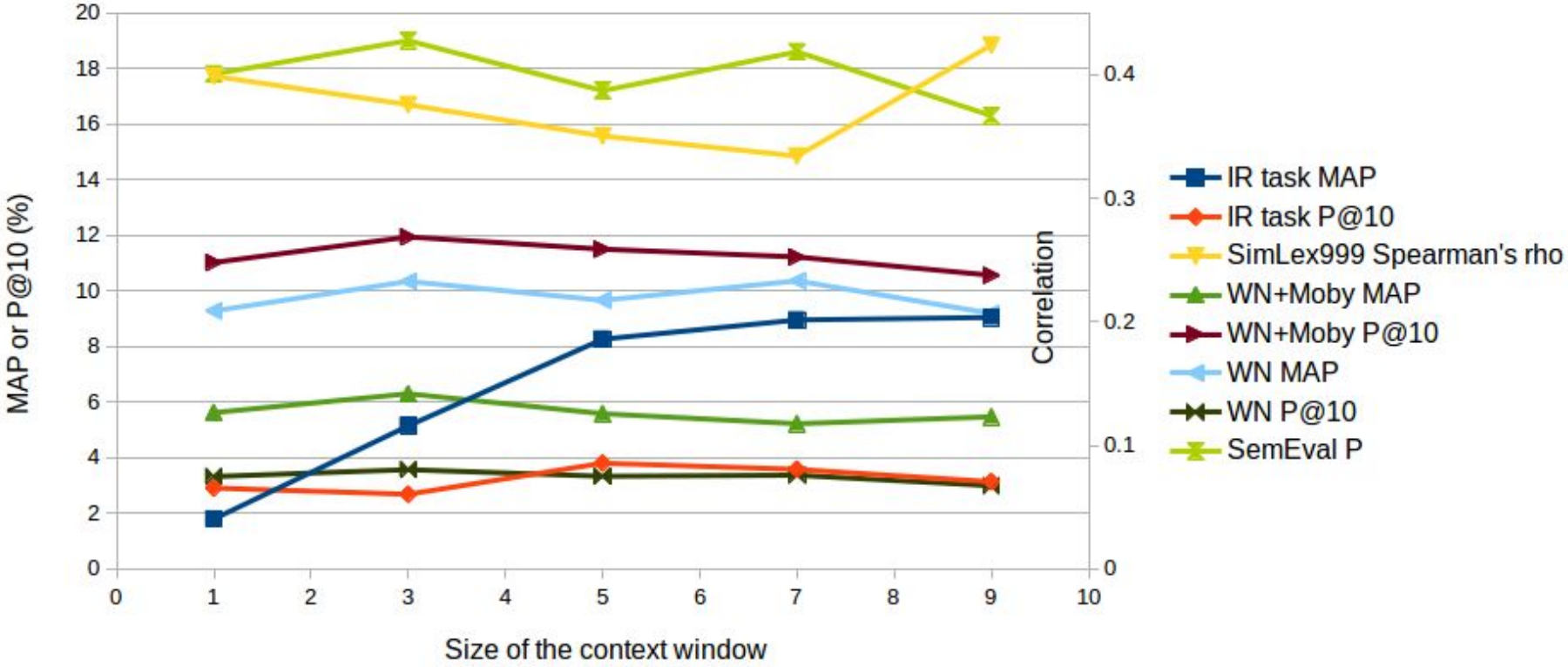
`sentence, abuse, detainee, guard, custody, defendant, inmate, prisoner`

☞ que se passe-t-il si on étend les requêtes qu'avec ces faux positifs ?

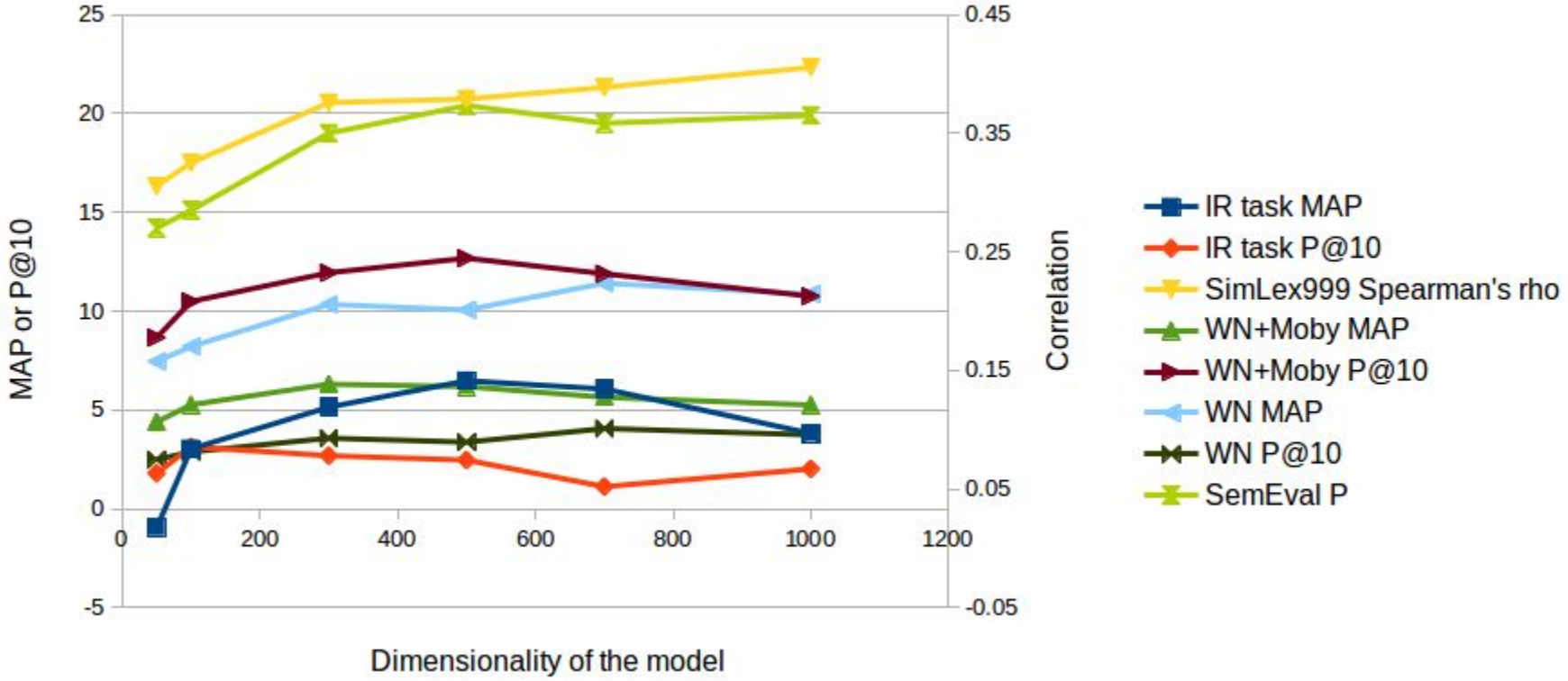
Ressources incomplètes ?



Un petit mot sur word2vec



Un petit mot sur word2vec



TAL pour la RI

Phonétique
- Sons -> mots

Morphologie
- formation des mots

Syntaxe
- formation d'énoncés

Sémantique
- sens des mots / énoncés

Pragmatique
- contexte, connaissance du monde

Connaissances, contextes

Vers une RI conceptuelle

- `Premier ministre du Canada // Justin Trudeau`
- `Losartan // hypertension`
- paraphrasage et ambiguité

Ressources

- DBpedia, graphe de connaissance, ontologies du domaine, annotation conceptuelle, règles d'inférence
- inclusion en RI : difficile, systèmes ad hoc

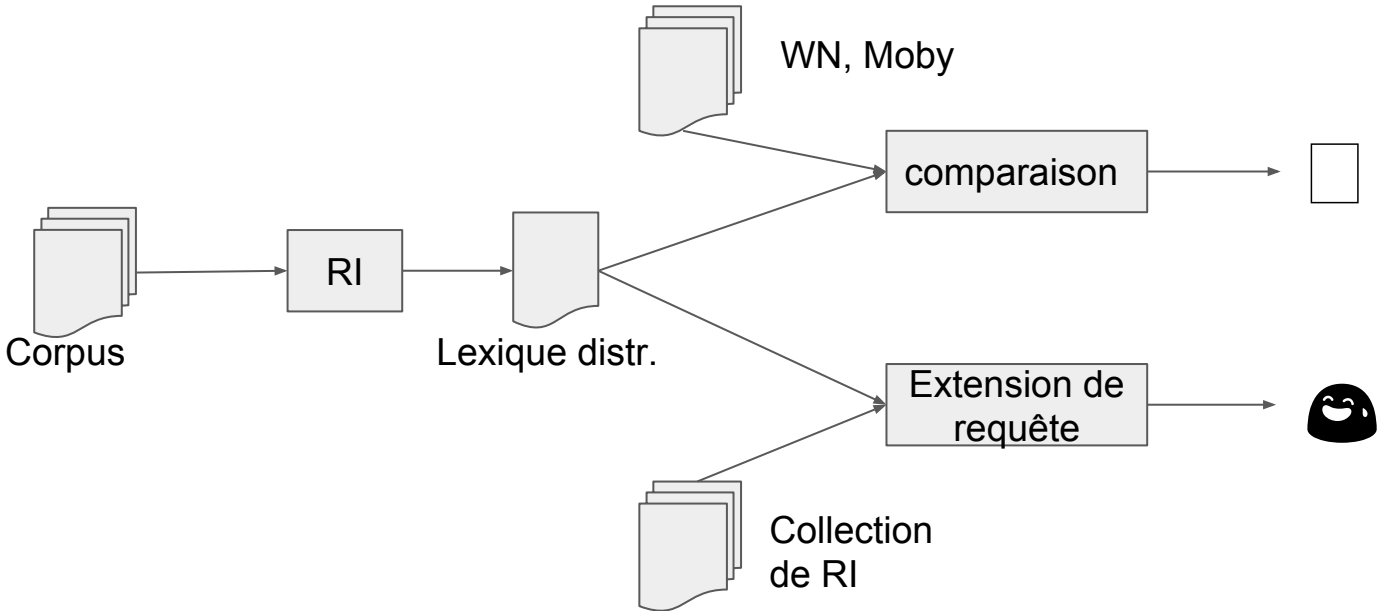
Résultats

- pas encore de gain évident [Zargayouna et al. 2015]

Conclusion

Ressources à tous les étages

- RI = ressources (outils) pour construire des lexiques distributionnels
- WN et Moby = ressources (données) d'évaluation
- RI = ressources (données + protocole) d'évaluation



TAL, RI...

Couplage productif mais délicat

- beaucoup de résultats mitigés dans la littérature
- morphologie, sémantique distributionnelle bénéfiques pour la RI
- évaluation par ressources externes ≠ évaluation par la tâche

Conditions de succès

- outils adaptés au contenu des documents
- capacité d'inclure les informations dans le processus
- utilisation peut être différente de celle d'origine

TAL, RI… et IA ?

Apprentissage artificiel

- représentation : *embeddings*, *deep learning* deplacent l'expertise
  - prise en compte de plusieurs phénomènes linguistiques
  - représentation de mots OK, repr. des textes plus difficiles, repr. textes + connaissances est un enjeu important
- apprentissage du système de RI
  - tâche supervisée (logs), metric learning, learning to rank

Sciences des données

- corpus scientifique commun RI, TAL, multimedia…
- outils communs

Pour aller plus loin

RI et TAL

- Moreau F., Sébillot P., Contributions des techniques du traitement automatique des langues à la recherche d'information, Rapport de recherche no 1690, IRISA, 2005.
- Traitement Automatique des Langues, numéro spécial recherche d'information, 56(3), J.-Y. Nie, V. Claveau
- Claveau V., Kijak E., « Thésaurus distributionnels pour la recherche d'information et vice-versa », Revue des Sciences et Technologies de l'Information - Série Document Numérique, 2015.

Semaine des technologies de la langue

15-18 mai 2018

Inria - IRISA Rennes

- CORIA
- TALN
- RJC
- Salon de l'innovation
- tutoriels
- Ateliers, hackathon