M2SOL023 - Dictionnaire et Néologismes

Le lexique vu côté informatique

Gaël Lejeune gael.lejeune@sorbonne-universite.fr

2024-2025

STIH EA 4509, Sorbonne Université

Rappels sur l'organisation générale

Modalités d'évaluation :

- 20 % Rendus en TD
- **30 % Réalisation d'un projet** (mini-mémoire + code) :
  - Présentation orale + dossier écrit
  - Oral de soutenance des mini-mémoires se fera le 15 mai (dernière séance)

Rappels sur l'organisation générale

Modalités d'évaluation :

- 20 % Rendus en TD
- **30 % Réalisation d'un projet** (mini-mémoire + code) :
  - Présentation orale + dossier écrit
  - Oral de soutenance des mini-mémoires se fera le 15 mai (dernière séance)
- 50% Examen final sur machine
- Analyse d'un problème, résolution par du code

Sujets des années passées

- Etude diachronique des expressions de souhaits dans les discours politiques
- L'évolution du langage à travers les paroles de musique
- Les marques de l'oralité sur Internet
- Analyse de sentiments dans les tweets
- Analogies phonétiques et sémantiques
- Analyse diachronique de mangas
- Analyse de la polarité d'avis Steam
- Comparaison de la couverture de dictionnaires
- Néologie ...

Sujets proposés cette année

Travail sur corpus pour des recherches sur le Lexique/les usages, en synchronie ou en diachronie.

Sujets proposés

- Compléments d'Objet Implicites
- Identification automatique de défigements
- Lexique et débats politiques
- Entités nommées spatiales en contexte bruité
- Versification et néologismes
- Le lexique autour des lieux de Paris
- La liberté d'expression dans la presse
- Comment caractériser la langue d'Annie Ernaux ?
- Terminologie dans le domaine de l'écologie : l'exemple de la dégradation des terres

Programme suite et fin

- Séance sur l'alignement (entre le mot et la phrase)
- Séance sur la néologie en diachronie
- Focus : Veille épidémiologique

Veille épidémiologique multilingue : Une approche parcimonieuse au grain caractère fondée sur le genre textuel

Introduction

Cadre général

Recherche d'information et multilinguisme

- La quantité de documents en ligne augmente
- **La part relative de l'anglais diminue**

Cadre général

Recherche d'information et multilinguisme

- La quantité de documents en ligne augmente
- **La part relative de l'anglais diminue**
- 10 langues représentent 85% des pages Web...
- Mais moins de 50% de la population mondiale

Cadre général

Recherche d'information et multilinguisme

- La quantité de documents en ligne augmente
- **La part relative de l'anglais diminue**
- 10 langues représentent 85% des pages Web...
- Mais moins de 50% de la population mondiale
- 50% de la population a donc accès à 15% des documents
- La démocratisation de l'Internet crée un rééquilibrage

Cadre général

Recherche d'information et multilinguisme

- La quantité de documents en ligne augmente
- **La part relative de l'anglais diminue**
- 10 langues représentent 85% des pages Web...
- Mais moins de 50% de la population mondiale
- 50% de la population a donc accès à 15% des documents
- La démocratisation de l'Internet crée un rééquilibrage

La part des **langues peu traitées** augmente

Plan

Introduction

La veille épidémiologique

- Analyser les articles de presse en ligne
- Détecter les foyers d'épidémies
- Surveiller leur évolution

Introduction

La veille épidémiologique

- Analyser les articles de presse en ligne
- Détecter les foyers d'épidémies
- Surveiller leur évolution

Les enjeux du multilinguisme dans la veille

- Augmenter la couverture
- Maîtriser le bruit
- Accélérer la détection

Introduction

La veille épidémiologique

- Analyser les articles de presse en ligne
- Détecter les foyers d'épidémies
- Surveiller leur évolution

Les enjeux du multilinguisme dans la veille

- Augmenter la couverture
- Maîtriser le bruit
- Accélérer la détection

Le traitement par l'humain est-il suffisant ?

- Délai de détection
- Coût important

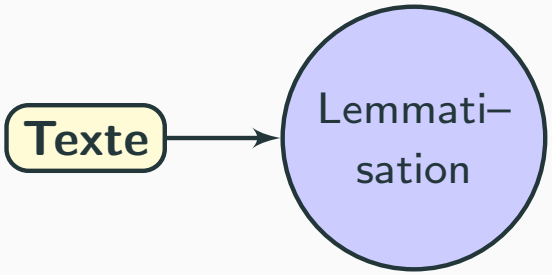
État de l'art (I)

Extraction d'information
Approche fondée sur une chaîne de traitement classique de TAL (Hobbs 93)

Texte

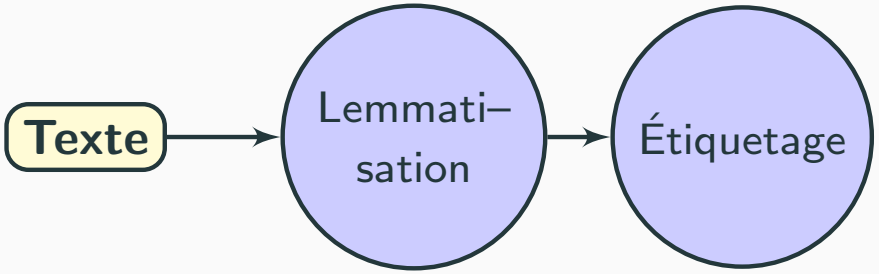
État de l'art (I)

Extraction d'information
Approche fondée sur une chaîne de traitement classique de TAL (Hobbs 93)



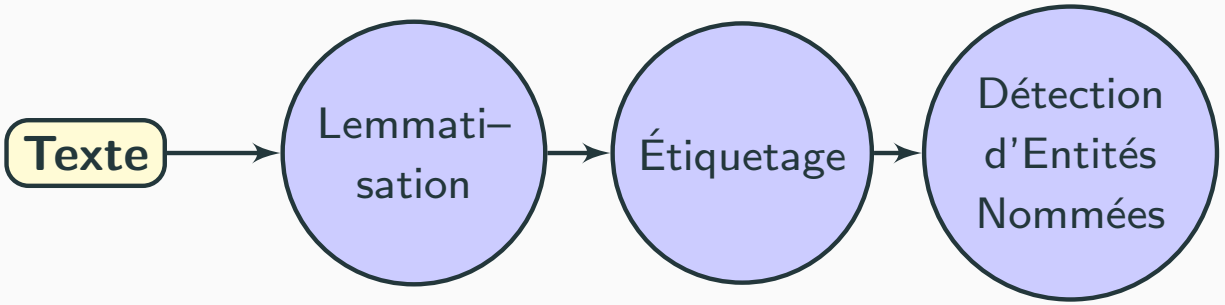
État de l'art (I)

Extraction d'information
Approche fondée sur une chaîne de traitement classique de TAL (Hobbs 93)



État de l'art (I)

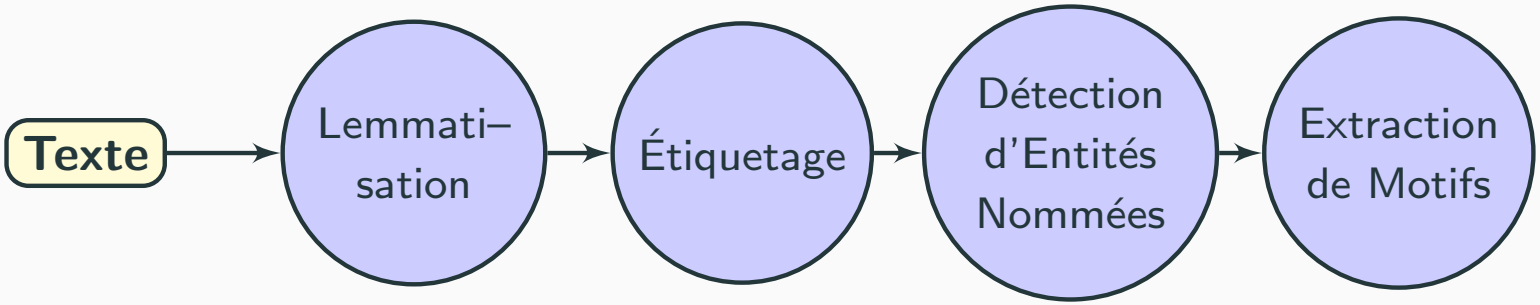
Extraction d'information
Approche fondée sur une chaîne de traitement classique de TAL (Hobbs 93)



État de l'art (I)

Extraction d'information

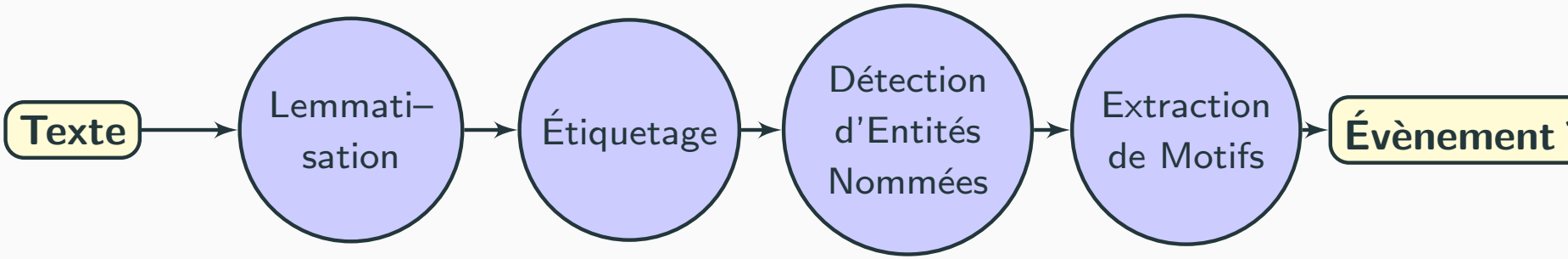
Approche fondée sur une chaîne de traitement classique de TAL (Hobbs 93)



État de l'art (I)

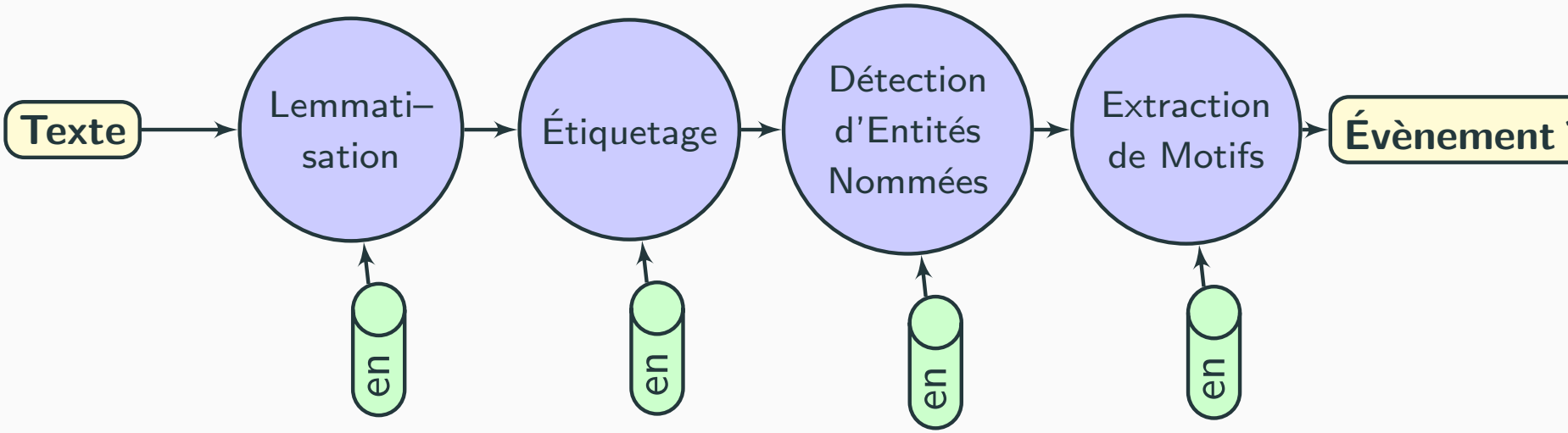
Extraction d'information

Approche fondée sur une chaîne de traitement classique de TAL (Hobbs 93)



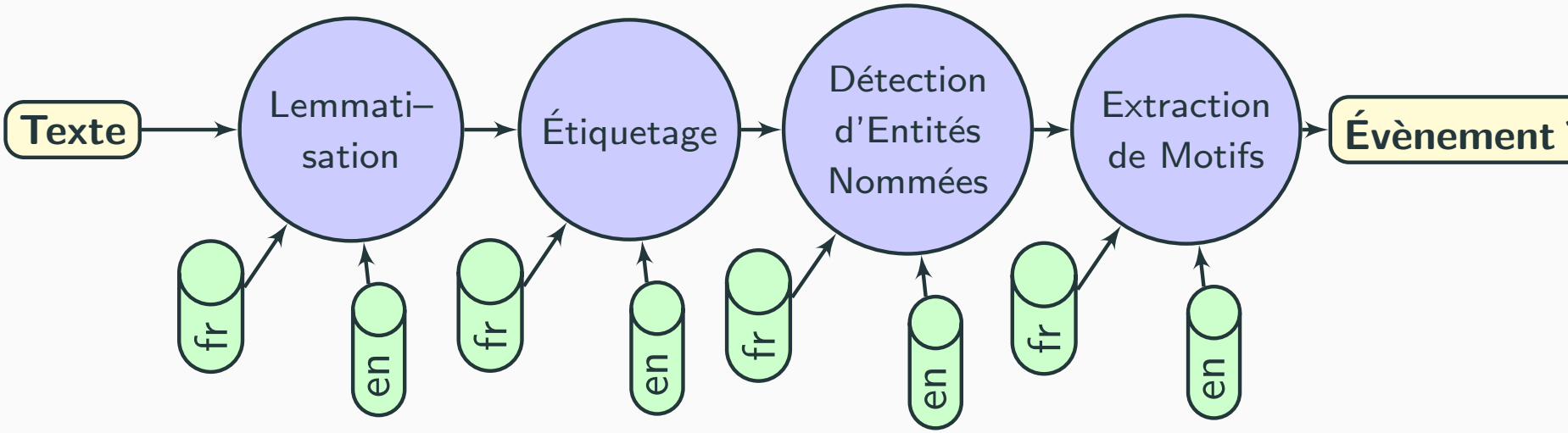
État de l'art (I)

Extraction d'information
Approche fondée sur une chaîne de traitement classique de TAL (Hobbs 93)



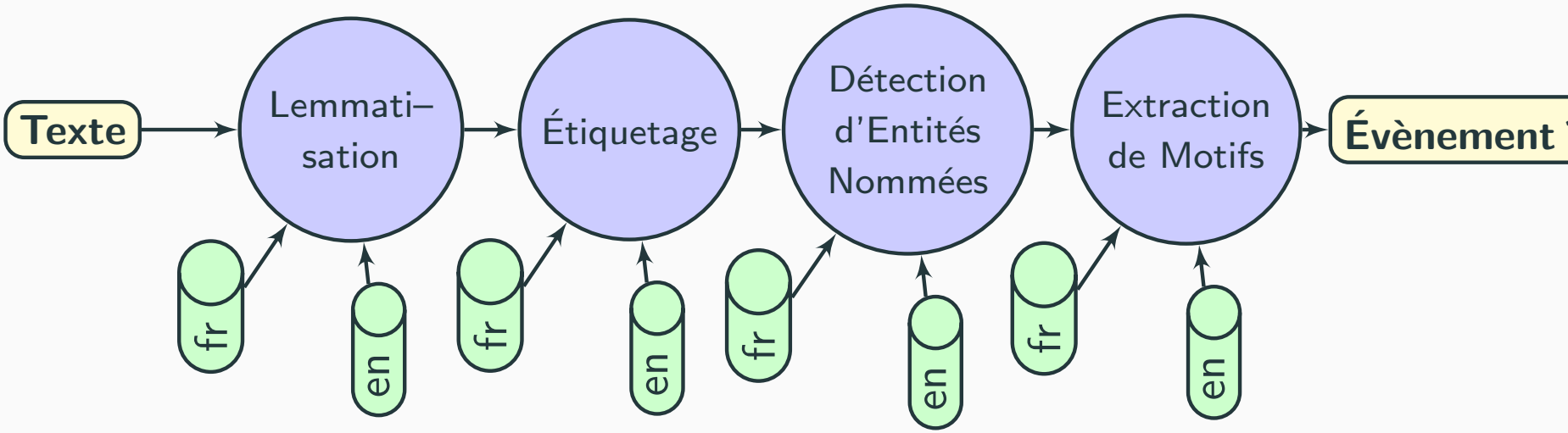
État de l'art (I)

Extraction d'information
Approche fondée sur une chaîne de traitement classique de TAL (Hobbs 93)



État de l'art (I)

Extraction d'information
Approche fondée sur une chaîne de traitement classique de TAL (Hobbs 93)



Une nouvelle langue ≡ nombreuses ressources

État de l'art (II)

Utiliser la traduction automatique

- Traduire les textes vers une langue déjà traitée (Piskorski-2011, Collier-2012)
- Traduire les motifs vers une nouvelle langue (Yangarber-2011)

État de l'art (II)

Utiliser la traduction automatique

- Traduire les textes vers une langue déjà traitée (Piskorski-2011, Collier-2012)
- Traduire les motifs vers une nouvelle langue (Yangarber-2011)

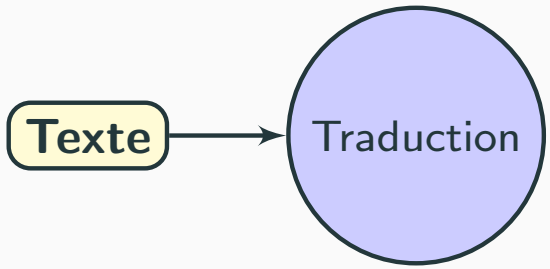
Texte

État de l'art (II)

Utiliser la traduction automatique

- Traduire les textes vers une langue déjà traitée (Piskorski-2011, Collier-2012)

- Traduire les motifs vers une nouvelle langue (Yangarber-2011)

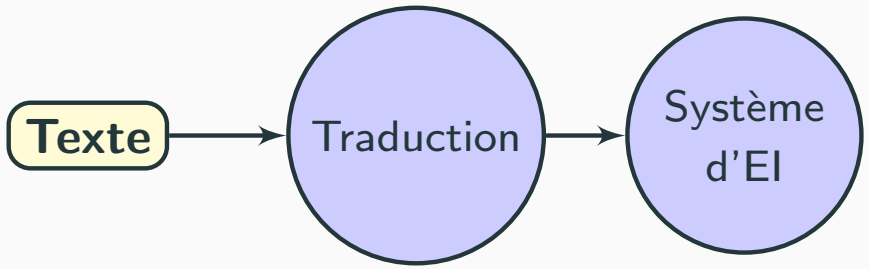


État de l'art (II)

Utiliser la traduction automatique

- Traduire les textes vers une langue déjà traitée (Piskorski-2011, Collier-2012)

- Traduire les motifs vers une nouvelle langue (Yangarber-2011)

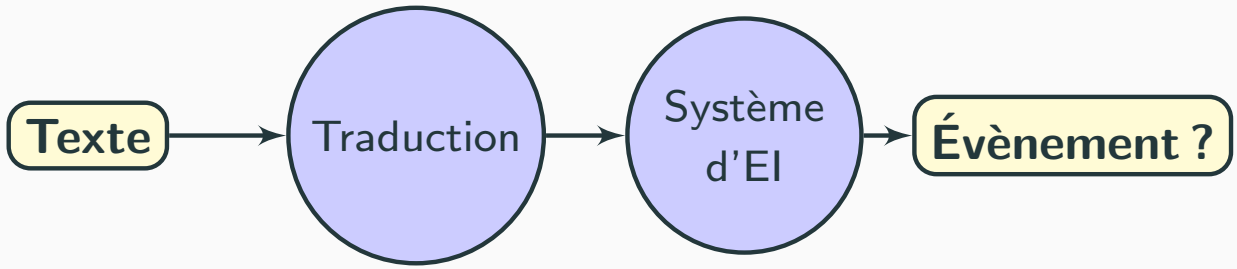


État de l'art (II)

Utiliser la traduction automatique

- Traduire les textes vers une langue déjà traitée (Piskorski-2011, Collier-2012)

- Traduire les motifs vers une nouvelle langue (Yangarber-2011)

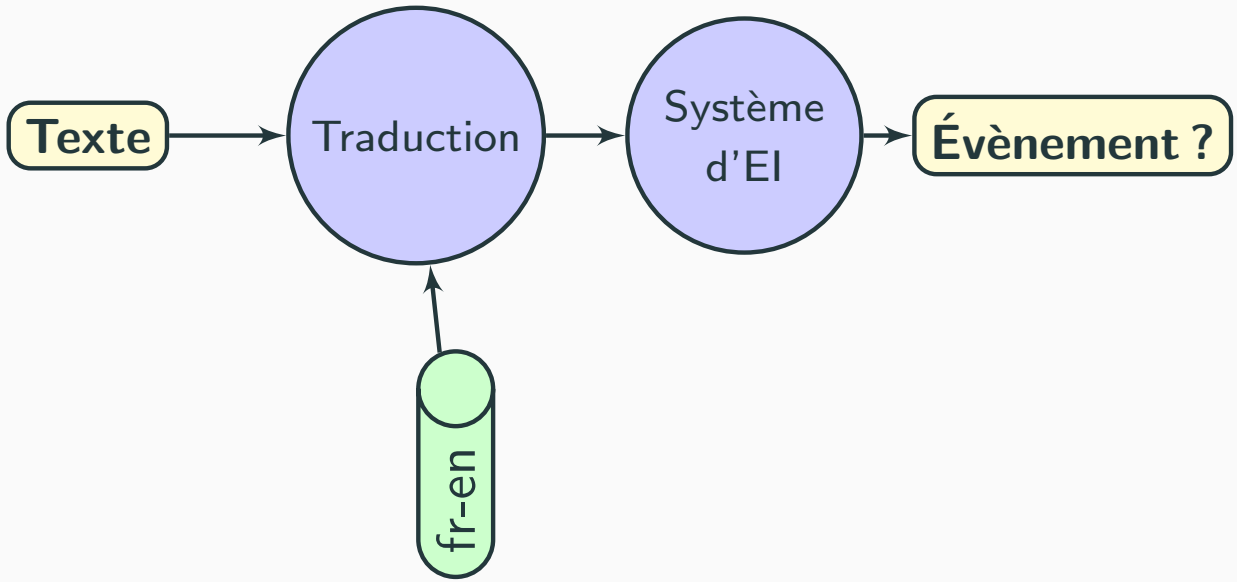


État de l'art (II)

Utiliser la traduction automatique

- Traduire les textes vers une langue déjà traitée (Piskorski-2011, Collier-2012)

- Traduire les motifs vers une nouvelle langue (Yangarber-2011)

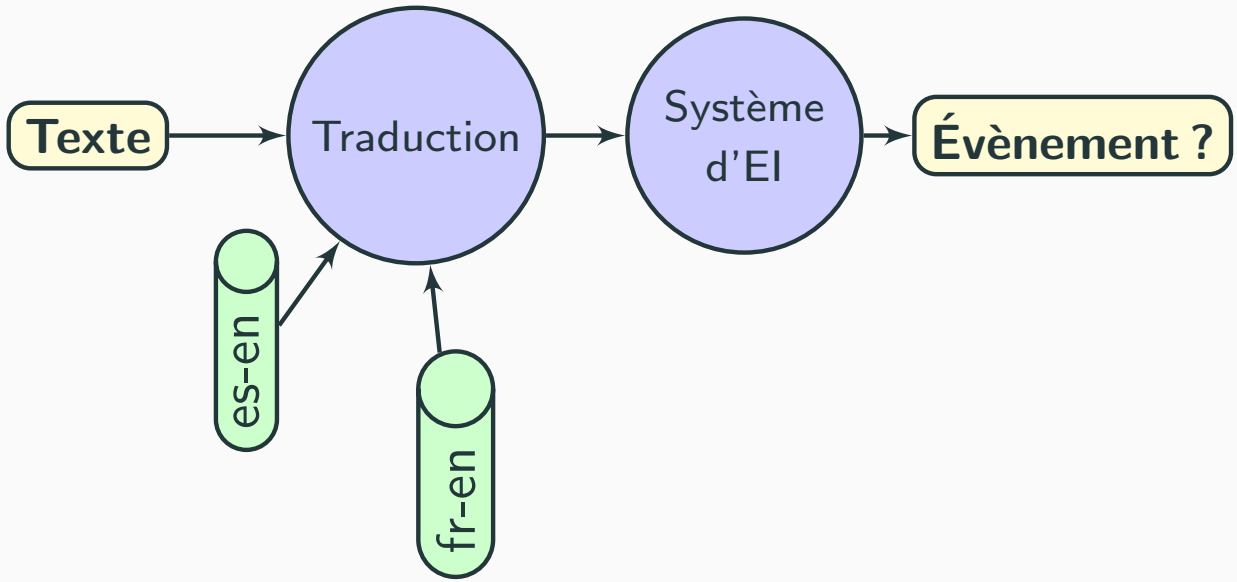


État de l'art (II)

Utiliser la traduction automatique

- Traduire les textes vers une langue déjà traitée (Piskorski-2011, Collier-2012)

- Traduire les motifs vers une nouvelle langue (Yangarber-2011)

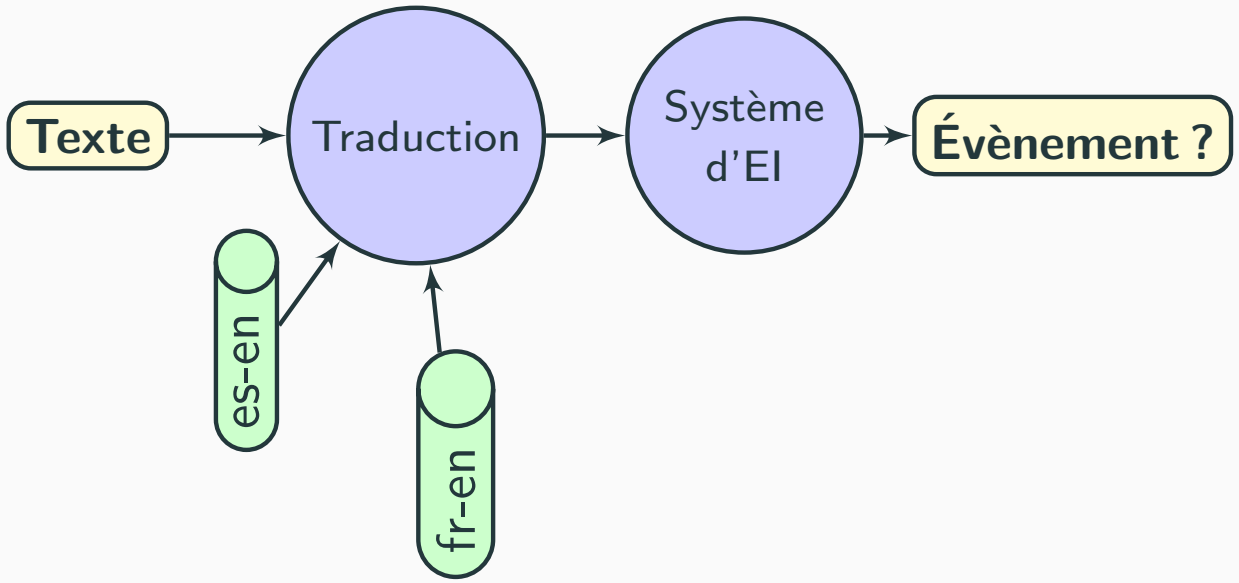


État de l'art (II)

Utiliser la traduction automatique

- Traduire les textes vers une langue déjà traitée (Piskorski-2011, Collier-2012)

- Traduire les motifs vers une nouvelle langue (Yangarber-2011)



Une nouvelle langue ≡ nouveau module de TA

Couverture

Langues (Nb. locuteurs 10^6)	Types de systèmes: Manuel	Semi-automatiques		Automatiques		
	ProMED	GPHIN	HealthMap	PULS	Biocaster	**DAnIEL**
Anglais (1,000)	X	X	X	X	X	X
Russe (277)	X	X	X	X	X	X
Espagnol (500)	X	X	X		X	X
Français (200)	X	X	X		X	X
Arabe (255)		X	X		X	X
Portugais (240)	X		X		X	X
Chinois (1,151)		**X**	X			X
Coréen (78)					X	X
Thaï (60)					X	X
Vietnamien (86)					X	X
Allemand (166)						**X**
Finnois (5)						**X**
Grec (13)						**X**
Italien (62)						**X**
Néerlandais (21)						**X**
Norvégien (5)						**X**
Polonais (46)						**X**
Suédois (8)						**X**
Tchèque (10)						**X**
Turc (75)						**X**

DAnIEL : approche multilingue

DAnIEL : approche multilingue

Constat : augmenter le nombre de langues couvertes est coûteux
Pour réaliser une couverture multilingue, il faut diminuer le **coût marginal** de traitement d'une nouvelle langue.

DAnIEL : approche multilingue

Constat : augmenter le nombre de langues couvertes est coûteux
Pour réaliser une couverture multilingue, il faut diminuer le **coût marginal** de traitement d'une nouvelle langue.

Objectif : augmenter la couverture linguistique et géographique

- Faciliter le travail des autorités sanitaires
- Traiter de grandes quantités d'articles :
  - *European Media Monitor*, 40000 articles/jour, 43 langues
- Détecter un évènement dès que possible (le premier article en langue vernaculaire)

DAnIEL : approche multilingue

DAnIEL
Data Analysis for Information Extraction in any Language

parcimonieux Limite la dépendance à des ressources externes

générique Utilise des propriétés du genre journalistique

multilingue Effectue une analyse au grain caractère

DAnIEL : approche multilingue

DAnIEL
Data Analysis for Information Extraction in any Language

parcimonieux Limite la dépendance à des ressources externes

générique Utilise des propriétés du genre journalistique

multilingue Effectue une analyse au grain caractère

Deux tâches pour comparer DAnIEL à l'état de l'art

- Classification de documents
- Extraction d'évènements épidémiologiques

DAnIEL : Approche multilingue

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." [. . .]

DAnIEL : Approche multilingue

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." [. . .]

Les plus longues chaînes de caractères répétées sont en **gras**

Approche multilingue

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." [. . .]

*Après filtrage par une liste de taille limitée, la paire (**maladie**-**lieu**) apparaît.*

DAnIEL : approche multilingue

Le russe

- Richesse de la morphologie
- Manque de modules de traitement

DAnIEL : approche multilingue

Le russe

- Richesse de la morphologie
- Manque de modules de traitement

Deux approches de l'état de l'art

- **PULS** : compléter une chaîne classique d'extraction d'information (Yangarber-2012)
- **Biocaster** : utiliser des modules de traduction automatique puis la chaîne de traitement de l'anglais (Collier-2011)

DAnIEL : approche multilingue

Онищенко призвал туристов к бдительности из-за птичьего гриппа

Эксперты не исключают вспышки вируса птичьего гриппа в КитаеРоссиян при выезде за рубеж просят проводить профилактику от А(H5N1)Грипп типа А(H1N1): профилактика и лечение. Справка

инфографика: Советы туристу, собирающемуся за границу

МОСКВА, 3 янв - РИА Новости. Роспотребнадзор рекомендует россиянам, отдыхающим в странах Юго-Восточной Азии, внимательнее следить за своим здоровьем в связи с регистрацией случая смерти от птичьего гриппа в Китае, сообщил РИА Новости во вторник главный государственный санитарный врач России Геннадий Онищенко.

"Уже в первые дни нового года, буквально первого, второго января, к сожалению, пришло сообщение, что на территории Китая гражданин умер от вируса птичьего гриппа. Поэтому эти настораживающие признаки, которые к нам доходят из регионов Юго-Восточной Азии, дают нам основание для дополнительной мобилизации", - сказал Онищенко.

В последний день 2011 года стало известно, что житель провинции Гуандун на юге Китая скончался от вируса птичьего гриппа. Это первый случай гибели людей от птичьего гриппа в Китае за последние месяцы.

По данным агентства Синьхуа, 39-летний водитель автобуса был госпитализирован 21 декабря в районе Баоань города Шэньчжэнь на границе со Специальным административным районом Сянган (Гонконг). Как рассказал врачам сам заболевший, в последнее время он не контактировал с живой птицей и не покидал пределов Шэньчжэня. Медики установили, что мужчина скончался от вируса гриппа H5N1.

Figure 1: Grippe aviaire en Chine

Une approche fondée sur le genre textuel

Utilisation du genre journalistique . . .
Les journalistes utilisent **la position** et **la répétition** pour transmettre leur message (principe du **moindre coût cognitif**).

Une approche fondée sur le genre textuel

Utilisation du genre journalistique . . .
Les journalistes utilisent **la position** et **la répétition** pour transmettre leur message (principe du **moindre coût cognitif**).

. . . dans une approche parcimonieuse . . .
Exploiter ces indices permet de **limiter les ressources nécessaires**.

Une approche fondée sur le genre textuel

Utilisation du genre journalistique . . .
Les journalistes utilisent **la position** et **la répétition** pour transmettre leur message (principe du **moindre coût cognitif**).

. . . dans une approche parcimonieuse . . .
Exploiter ces indices permet de **limiter les ressources nécessaires**.

. . . au grain caractère.
Analyser les textes tels quels **sans analyse locale**.

Présentation générale de l'algorithme

Relevant_content
En ressource : noms de maladies collectés automatiquement sur Wikipedia

En entrée : l'article de presse

1. Segmenter l'article en zones d'intérêt (**positions**)
2. Rechercher les sous-chaînes de caractères communes entre ces zones et chacun des noms de maladie de la base (**répétitions**)

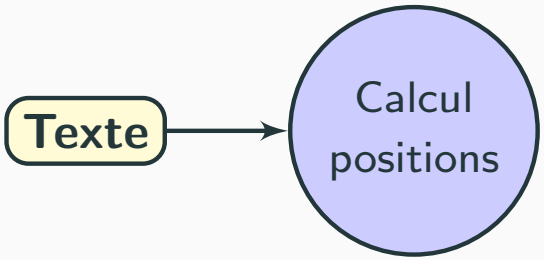
Texte

Présentation générale de l'algorithme

Relevant_content
En ressource : noms de maladies collectés automatiquement sur Wikipedia

En entrée : l'article de presse

1. Segmenter l'article en zones d'intérêt (**positions**)
2. Rechercher les sous-chaînes de caractères communes entre ces zones et chacun des noms de maladie de la base (**répétitions**)

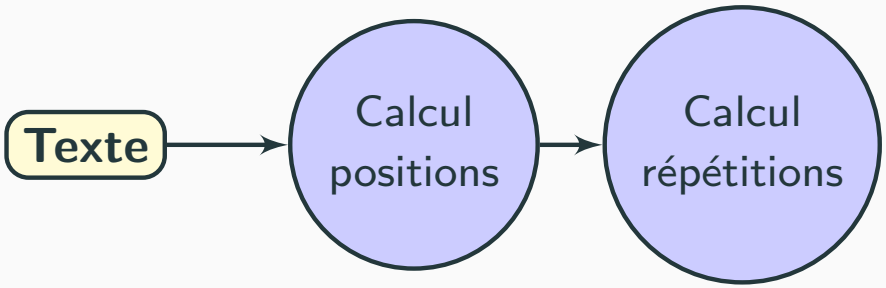


Présentation générale de l'algorithme

Relevant_content
En ressource : noms de maladies collectés automatiquement sur Wikipedia

En entrée : l'article de presse

1. Segmenter l'article en zones d'intérêt (**positions**)
2. Rechercher les sous-chaînes de caractères communes entre ces zones et chacun des noms de maladie de la base (**répétitions**)

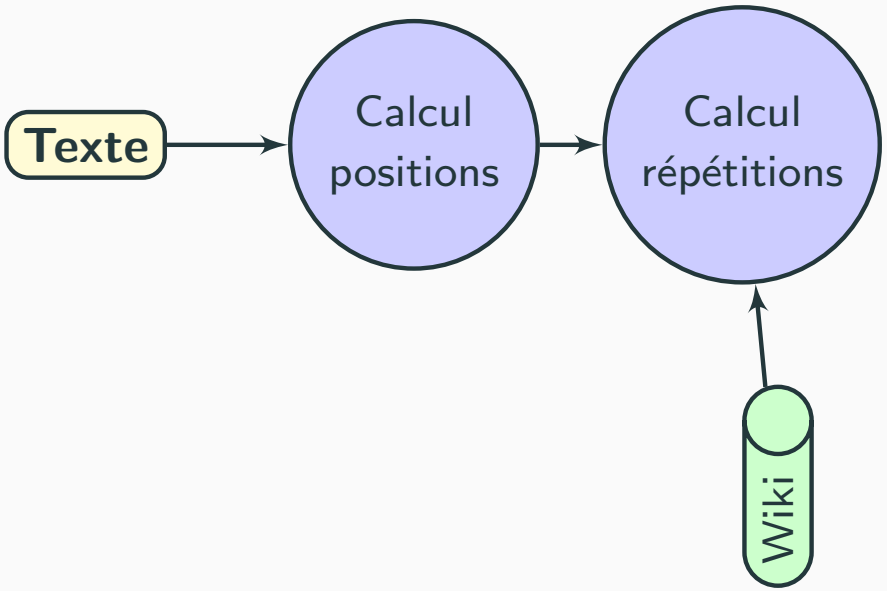


Présentation générale de l'algorithme

Relevant_content
En ressource : noms de maladies collectés automatiquement sur Wikipedia

En entrée : l'article de presse

1. Segmenter l'article en zones d'intérêt (**positions**)
2. Rechercher les sous-chaînes de caractères communes entre ces zones et chacun des noms de maladie de la base (**répétitions**)

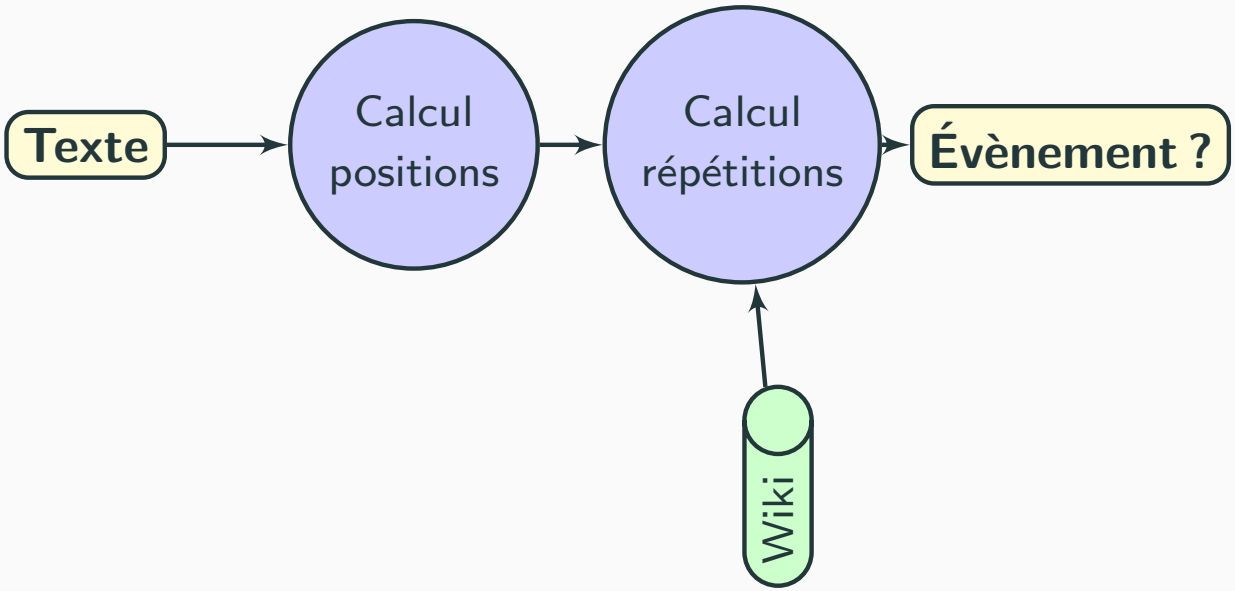


Présentation générale de l'algorithme

Relevant_content
En ressource : noms de maladies collectés automatiquement sur Wikipedia

En entrée : l'article de presse

1. Segmenter l'article en zones d'intérêt (**positions**)
2. Rechercher les sous-chaînes de caractères communes entre ces zones et chacun des noms de maladie de la base (**répétitions**)

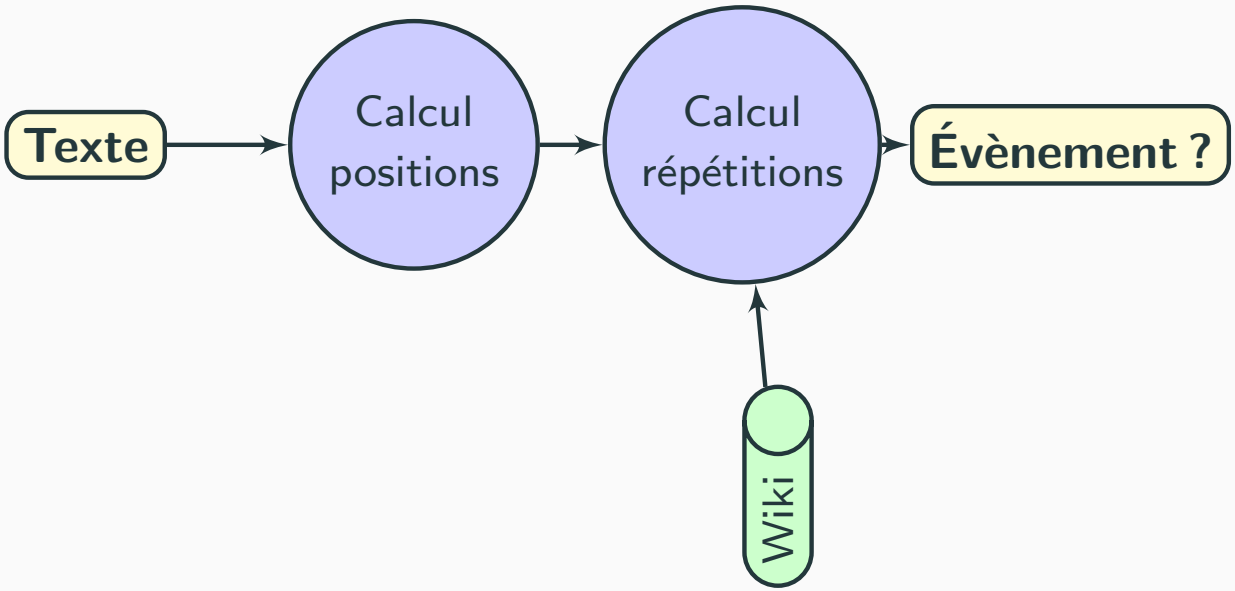


Présentation générale de l'algorithme

Relevant_content
En ressource : noms de maladies collectés automatiquement sur Wikipedia

En entrée : l'article de presse

1. Segmenter l'article en zones d'intérêt (**positions**)
2. Rechercher les sous-chaînes de caractères communes entre ces zones et chacun des noms de maladie de la base (**répétitions**)



Détermination des zones d'intérêt

Positions et répétitions
3 types d'articles journalistiques en fonction de la longueur :

- Court : toutes les répétitions sont intéressantes
- Moyen : on compare la tête (titre + chapeau) et le corps (reste du texte)
- Long : on ne considère que la tête et le pied (dernier paragraphe)

Segmentation empirique

Type d'article	Nb. paragraphes	Segments comparés
Court	3 et moins	Tous
Classique	4 à 10	Tête et corps
Long	11 et plus	Tête et pied

Extraction des motifs

Relevant_content en détails

1. Comparer les zones d'intérêt et les éléments du lexique
2. Extraire les chaînes répétées maximales (s)
3. Conserver celles présentes :
   - dans 2 zones d'intérêts
   - et dans un élément du lexique (m)
4. On considère qu'il y a un évènement épidémiologique lorsque $\frac{len(s)}{len(m)} > \theta$

Extraction des motifs

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. - Sugeruję, 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.

Extraction des motifs

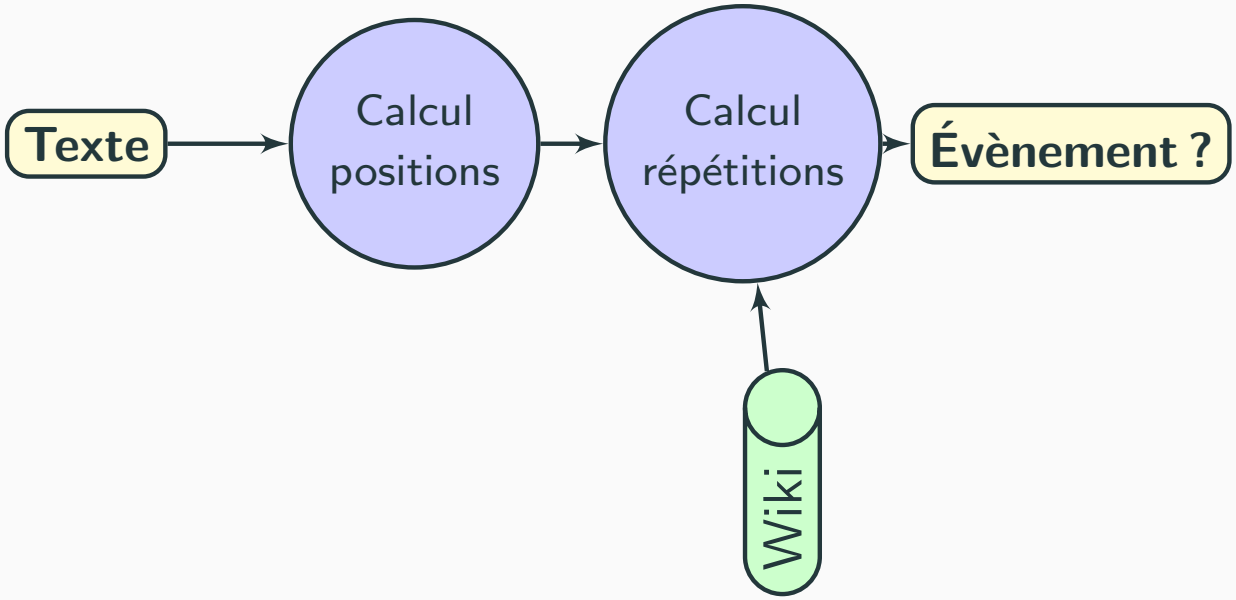
Quels motifs sont pertinents ?

On considère qu'il y a un évènement épidémiologique lorsque $\frac{len(s)}{len(m)} > \theta$

Table 1: Motifs extraits après filtrage positionnel et lexical classés par taille décroissante

$\frac{len(s)}{len(m)}$	Adaptation	sous-chaîne	terme polonais	traduction	positions
0.8	+	deng	denga	dengue	[1,4]
0.75	+	iła	kiła	syphilis	[1,3]
0.5	-	ła	kiła	syphilis	[1,2,3,5]
0.5	-	ró	róża	érysipèle	[2,5]
0.5	-	dr	odra	rougeole	[0,2]
0.44	-	rowi	norowirus	norovirus	[0,2]
0.4	-	ry	grypa	grippe	[1,5]
0.4	-	ng	denga	dengue	[1,4]

DAnIEL : approche multilingue



Une seule ressource externe impliquée

Évaluation comparative

Évaluation

Mesurer la plus-value offerte par DAnIEL par rapport à ProMED
Peut-on améliorer la vitesse de détection d'évènements épidémiologiques ?

Jeux de données
Période : octobre 2011 à février 2012

ProMED 1548 comptes-rendus structurés avec une Paire Maladie-Lieu (PML)

DAnIEL 1671 documents pertinents parmi 26091 articles de presse (17 langues)

Intersection : 167 évènements extraits par les deux systèmes (PML uniques)

Évaluation

Table 2: Zone géographique des évènements détectés en premier par DAnIEL

	Langues	Nb. Premières alertes
Amérique du Nord	en,es	4/26
Amérique du Sud et Centrale	en,es,pt	9/24
Afrique du Nord	**ar**,fr	3/8
Afrique sub-saharienne	fr	3/14
France, Portugal, Espagne, RU	en,es,fr,**nl**,pt	12/43
Reste de l'Europe	**cs**,**de**,**el**,**fi**,fr,**it**,**sv**	**12/19**
Russie & Ukraine	**pl**,ru	6/10
Chine & Inde	**zh**,en	3/8
Reste de l'Asie	**zh**,ru	9/15
Ensemble	15	61/167 (37%)

Évaluation

Table 2: Zone géographique des évènements détectés en premier par DAnIEL

	Langues	Nb. Premières alertes
Amérique du Nord	en,es	4/26
Amérique du Sud et Centrale	en,es,pt	9/24
Afrique du Nord	**ar**,fr	3/8
Afrique sub-saharienne	fr	3/14
France, Portugal, Espagne, RU	en,es,fr,**nl**,pt	12/43
Reste de l'Europe	**cs**,**de**,**el**,**fi**,fr,**it**,**sv**	**12/19**
Russie & Ukraine	**pl**,ru	6/10
Chine & Inde	**zh**,en	3/8
Reste de l'Asie	**zh**,ru	9/15
Ensemble	15	61/167 (37%)

Langue vernaculaire : **compléter la couverture** et d'**accélérer la détection**, notamment pour des **langues peu dotées**.

Évaluation

Amélioration des résultats

- Augmentation de la couverture
- Délai de détection minimisé
- Importance de la langue vernaculaire
- Approche efficace pour des tâches nécessitant un rappel élevé

Évaluation

Amélioration des résultats

- Augmentation de la couverture
- Délai de détection minimisé
- Importance de la langue vernaculaire
- Approche efficace pour des tâches nécessitant un rappel élevé

Qu'en est-il de la précision ?

- Validité des alertes émises par DAnIEL mais pas par ProMED ?
- Bruit généré par DAnIEL ?

Évaluation qualitative

Évaluation

Constitution
Le grec ,le polonais et le russe (morphologie riche), le chinois (système d'écriture différent) et l'anglais (pour la comparaison)

- Documents annotés par des locuteurs natifs
- Catégorie santé : *Google News* (chinois, russe et anglais), journaux (grec et polonais)
- Période : octobre 2011 à janvier 2012
- Cible : la Paire Maladie-Lieu (PML)
- Corpus rendu disponible

Évaluation qualitative

- Vérifier la validité des alertes émises par DAnIEL
- Caractériser le bruit qu'il génère

Évaluation

Description
Ressource : base de noms de maladie (termes) de *Wikipedia*
Principe : le document traité est pertinent si

B1 **présence** d'un terme de la base

B2 **répétition** d'un terme de la base

B3 **répétition** d'un terme de la base à des **positions** clés

Métriques

Succès Vrais Positifs (VP) et Vrais Négatifs (VN)

Échecs Faux Positifs (FP) et Faux Négatifs (FN)

Évaluation

Description
Ressource : base de noms de maladie (termes) de *Wikipedia*
Principe : le document traité est pertinent si

B1 **présence** d'un terme de la base

B2 **répétition** d'un terme de la base

B3 **répétition** d'un terme de la base à des **positions** clés

Métriques

Succès Vrais Positifs (VP) et Vrais Négatifs (VN)

Échecs Faux Positifs (FP) et Faux Négatifs (FN)

Rappel (R) $\frac{VP}{VP+FN}$

Précision (P) $\frac{VP}{VP+FP}$

Mesure-F (F) $\frac{(1+\beta^2)*(P*R)}{(\beta^2*P)+R}$

Évaluation

Description
Ressource : base de noms de maladie (termes) de *Wikipedia*

Principe : le document traité est pertinent si

B1 **présence** d'un terme de la base

B2 **répétition** d'un terme de la base

B3 **répétition** d'un terme de la base à des **positions** clés

		anglais	chinois	grec	polonais	russe	ensemble
Baseline 1 (B1)	Mesure-F_1	0,47	0,64	0,57	0,54	0,71	**0,57**
présence	Mesure-F_2	0,69	0,82	0,76	0,71	0,80	0,74
Baseline 2 (B2)	Mesure-F_1	0,59	0,86	0,69	0,55	0,77	0,68
répétition	Mesure-F_2	0,75	0,94	0,8	0,58	0,78	0,76
Baseline 3 (B3)	Mesure-F_1	0,67	0,89	0,82	0,43	0,76	**0,71**
répétition+posit.	Mesure-F_2	0,69	0,95	0,88	0,37	0,76	**0,72**

Table 3: Évaluation de trois *baselines* : F_1-mesure et F_2-mesure

Le seuil θ

Quels motifs sont pertinents ?

On considère qu'il y a un évènement épidémiologique lorsque $\frac{len(s)}{len(m)} > \theta$

Table 4: Motifs extraits après filtrage positionnel et lexical classés par taille décroissante

$\frac{len(s)}{len(m)}$	Adaptation	sous-chaîne	terme polonais	traduction	positions
0.8	+	deng	denga	dengue	[1,4]
0.75	+	iła	kiła	syphilis	[1,3]
0.5	-	ła	kiła	syphilis	[1,2,3,5]
0.5	-	ró	róża	érysipèle	[2,5]
0.5	-	dr	odra	rougeole	[0,2]
0.44	-	rowi	norowirus	norovirus	[0,2]
0.4	-	ry	grypa	grippe	[1,5]
0.4	-	ng	denga	dengue	[1,4]

Évaluation

	anglais	chinois	grec	polonais	russe
θ	0.8	0.75	0.75	0.76	0.85
Précision	0.70	0.84	0.68	0.65	0.88
Rappel	0.89	1.0	0.96	0.87	0.86
Mesure F_1	0.78	0.91	**0.80**	0.77	0.87
Mesure F_2	0.84	0.96	**0.89**	0.86	0.87

Table 5: Qualité pour chaque langue du filtrage des documents pour le ratio optimal (θ)

Évaluation

	anglais	chinois	grec	polonais	russe
θ	0.8	0.75	0.75	0.76	0.85
Précision	0.70	0.84	0.68	0.65	0.88
Rappel	0.89	1.0	0.96	0.87	0.86
Mesure F_1	0.78	0.91	**0.80**	0.77	0.87
Mesure F_2	0.84	0.96	**0.89**	0.86	0.87

Table 5: Qualité pour chaque langue du filtrage des documents pour le ratio optimal (θ)

	Ensemble du corpus	
θ	$\theta = 0.8$	θ optimal par langue
Précision	0.69	0.75
Rappel	0.89	0.91
Mesure F_1	0.78	0.82
Mesure F_2	**0.84**	**0.88**

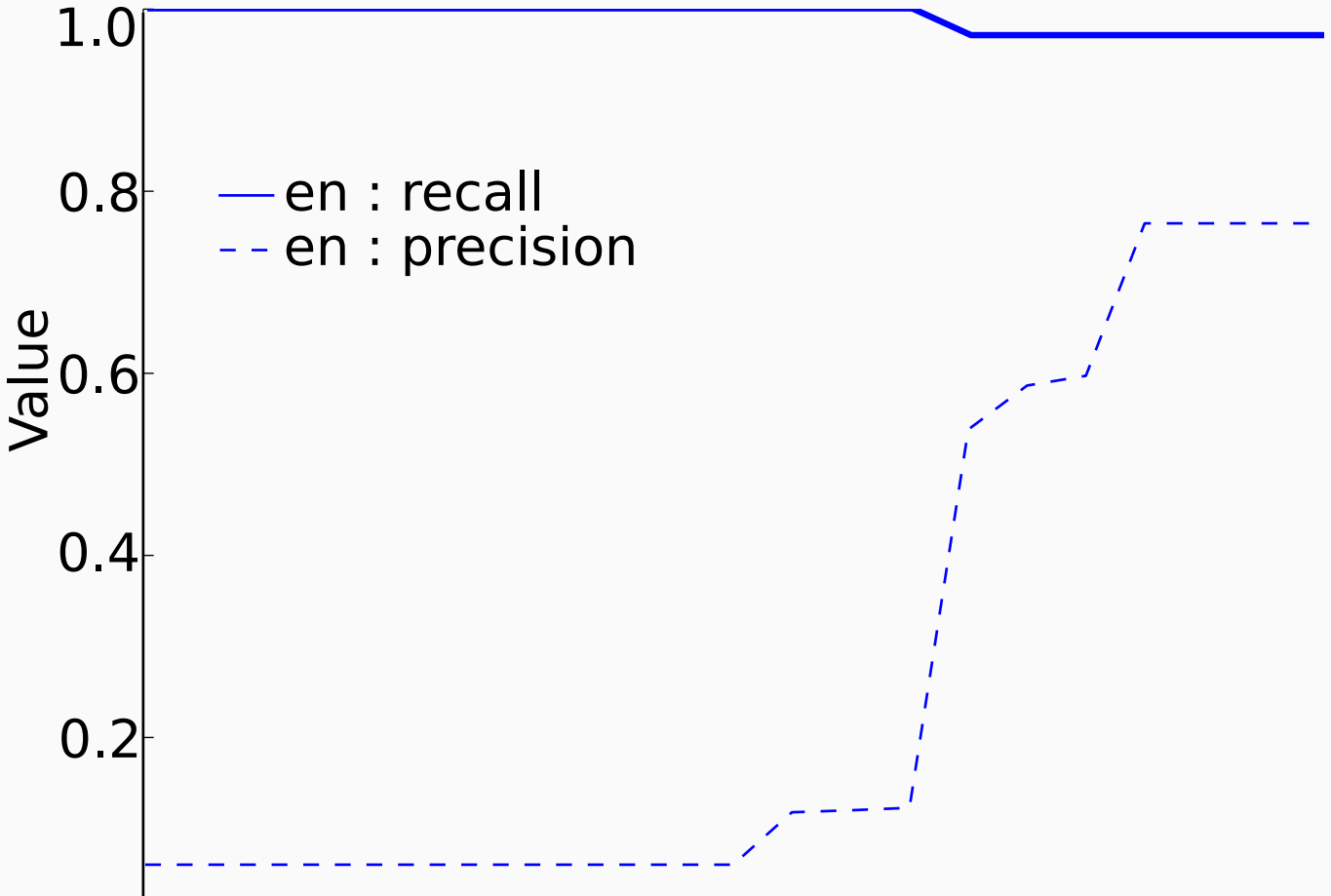
Table 6: Qualité pour l'ensemble des langues du filtrage des documents pour le ratio générique (0.8) et le ratio optimal

.3

(a)

b

Figure 2: Corpus anglais

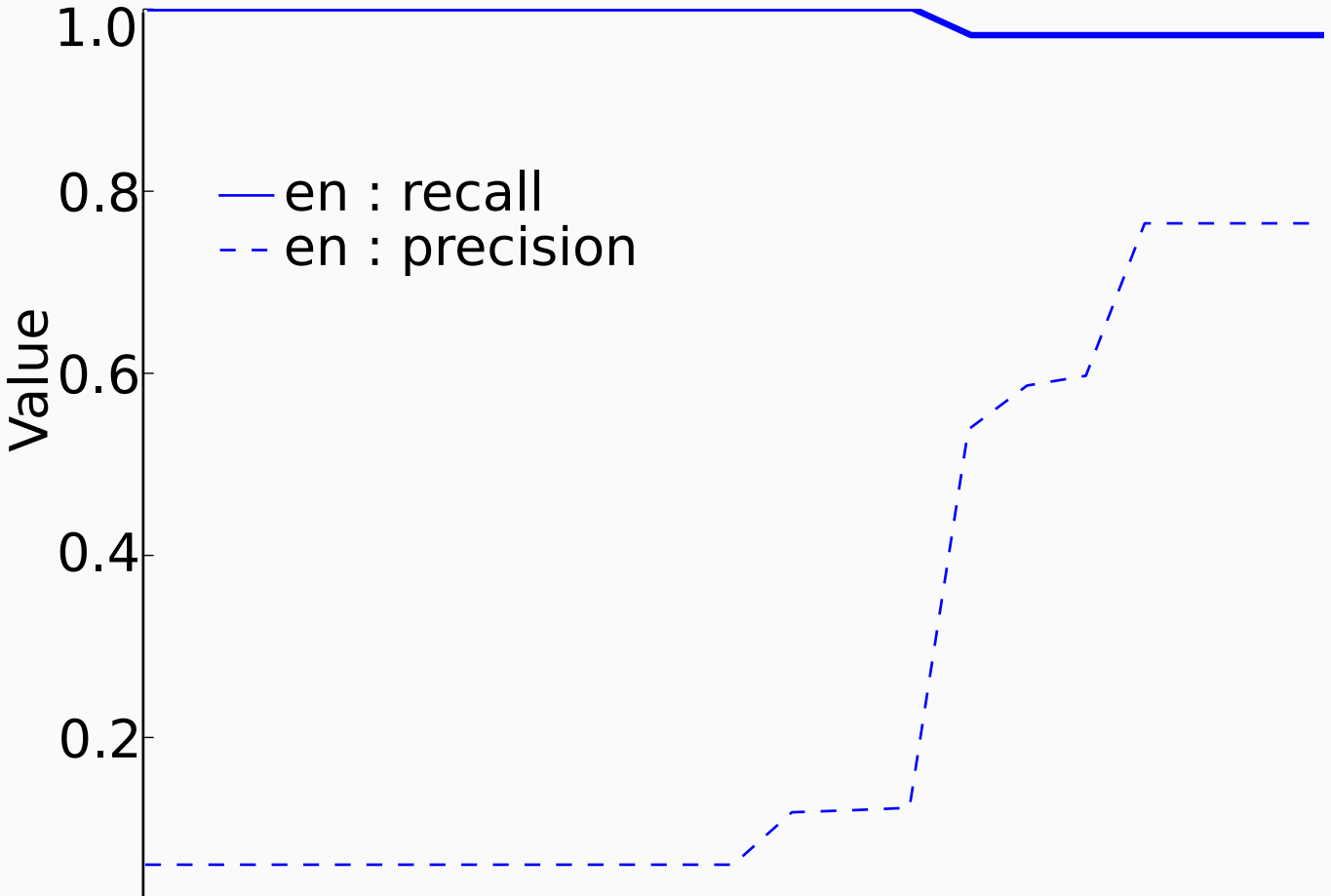


.3

(a)

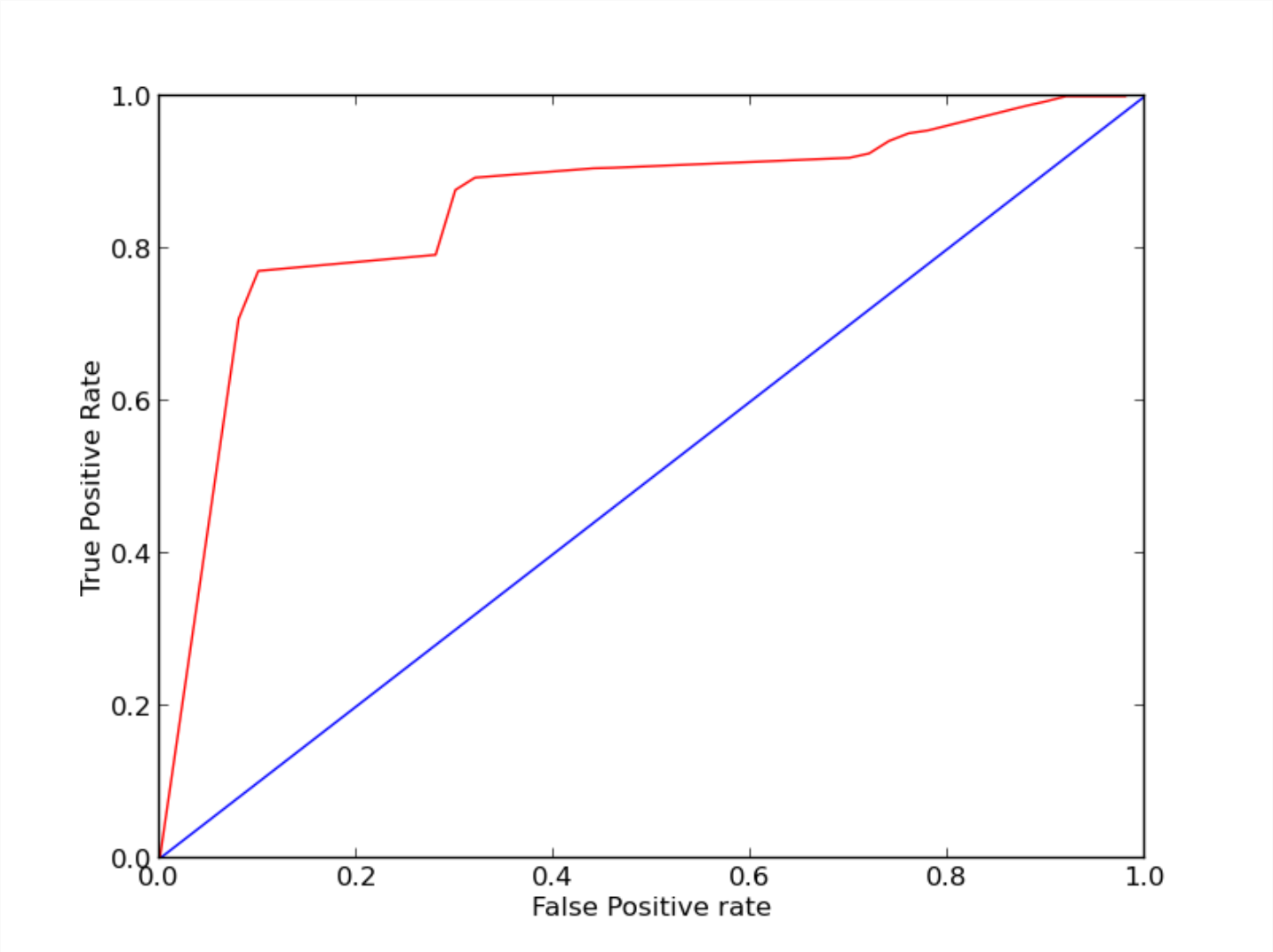
b

Figure 2: Corpus anglais



Évaluation

Figure 7: Courbe ROC sur notre échantillon de 2000 documents en 5 langues, aire sous la courbe : 0.86



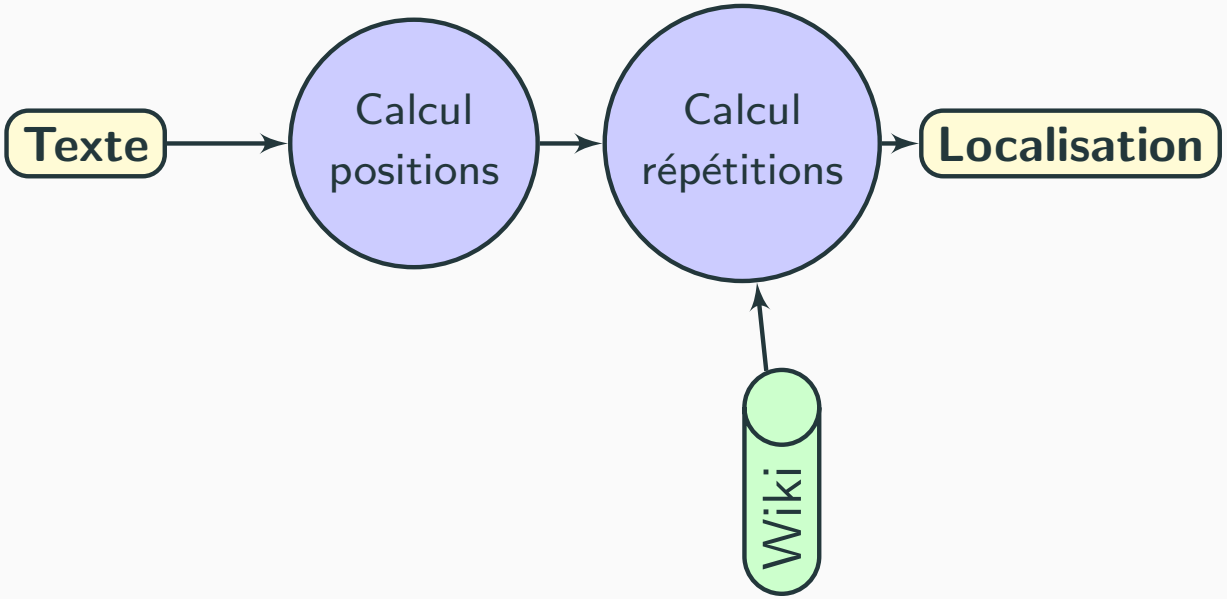
Évaluation

L'algorithme relevant_content est réutilisé
Si aucun nom de lieu connu n'est détecté (**pas de localisation explicite**), la localisation de l'évènement est celle de la source (**localisation implicite**).

Texte

Évaluation

L'algorithme relevant_content est réutilisé
Si aucun nom de lieu connu n'est détecté (**pas de localisation explicite**), la localisation de l'évènement est celle de la source (**localisation implicite**).



Wiki ≡ la liste de lieux

Évaluation

Les évènements sont-ils bien localisés ?

- La localisation explicite est-elle fiable ?
- L'hypothèse de la localisation implicite est-elle valide ?

Évaluation

Les évènements sont-ils bien localisés ?

- La localisation explicite est-elle fiable ?
- L'hypothèse de la localisation implicite est-elle valide ?

Localisation efficace à moindre coût

Table 7: Évaluation des règles de localisation

	anglais	grec	polonais	russe
Nombre d'évènements détectés	143	188	213	230
Pas de localisation explicite trouvée	46	33	35	51
Localisation = source	37	27	29	40
Localisation ≠ source	9	6	6	11
Taux global d'erreur	**12,2%**	3%	2,8%	4,8%

Évaluation

Pourquoi la précision est-elle notablement inférieure au rappel ?
Test sur les **évènements extraits par DAnIEL** à partir de **6000 documents** (grec, polonais et russe). On demande aux **annotateurs de juger si l'évènement est** :

- Très pertinent (nouveaux évènements, mise à jour ...)
- Assez pertinent (article de synthèse, campagnes de vaccination ...)
- Non pertinent (historique, secondaire, erroné ...)

Évaluation

Les limites de la classification binaire

	grec	polonais	russe
Très pertinent	82.7%	85.4%	82.3%
Assez pertinent	**11.9%**	8.0%	11.5%
Non pertinent	**5.4%**	6.6%	6.2%

Bilan

- Peu d'évènements réellement non-pertinents
- Difficulté de trancher pour certains articles
- Accord inter-annotateurs (Fleiss Kappa) **0.78 avec 2 classes** et **0.86 avec 3 classes**

Conclusion et perspectives

Conclusion

DAnIEL
Data Analysis for Information Extraction in any Language

parcimonieux Limite la dépendance à des ressources externes

générique Utilise des propriétés du genre journalistique

multilingue Effectue une analyse au grain caractère

Une approche économe et efficace

- Évaluation sur un corpus de plus de 2000 documents annotés
- 52 langues traitées, 2.6 ko de ressources par langue
- Plus-value sur les langues peu dotées
- 8000 documents/minute (PC portable 2 cœurs, 2.4GHz, 2Go RAM) en Python
- `https://daniel.greyc.fr` : corpus, résultats, test de DAnIEL

Perspectives

Améliorer l'approche

- Évaluer vis-à-vis d'autres systèmes

Perspectives

Améliorer l'approche

- Évaluer vis-à-vis d'autres systèmes
- Calculer les positions remarquables

Perspectives

Améliorer l'approche

- Évaluer vis-à-vis d'autres systèmes
- Calculer les positions remarquables
- Acquérir de la terminologie

Perspectives

Améliorer l'approche

- Évaluer vis-à-vis d'autres systèmes
- Calculer les positions remarquables
- Acquérir de la terminologie
- Relier des évènements

Perspectives

Améliorer l'approche

- Évaluer vis-à-vis d'autres systèmes
- Calculer les positions remarquables
- Acquérir de la terminologie
- Relier des évènements

Changer de variable

Perspectives

Améliorer l'approche

- Évaluer vis-à-vis d'autres systèmes
- Calculer les positions remarquables
- Acquérir de la terminologie
- Relier des évènements

Changer de variable

- La tâche : évaluation du risque, prévision d'évènements...

Perspectives

Améliorer l'approche

- Évaluer vis-à-vis d'autres systèmes
- Calculer les positions remarquables
- Acquérir de la terminologie
- Relier des évènements

Changer de variable

- La tâche : évaluation du risque, prévision d'évènements...
- Le domaine : veille économique, veille stratégique...

Perspectives

Améliorer l'approche

- Évaluer vis-à-vis d'autres systèmes
- Calculer les positions remarquables
- Acquérir de la terminologie
- Relier des évènements

Changer de variable

- La tâche : évaluation du risque, prévision d'évènements...
- Le domaine : veille économique, veille stratégique...
- Le genre : articles académiques, microblogs...