Vers une extraction automatique des événements dans les textes

Description d'un lexique pondéré des noms d'événements

Béatrice ARNULPHY Xavier TANNIER Anne VILNAT
{Beatrice.Arnulphy, Xavier.Tannier, Anne.Vilnat}@limsi.fr

Journée ATALA
Reconnaissance d'Entités Nommées -
Nouvelles Frontières & Nouvelles Approches
20 juin 2011

- Événements nominaux / EN Événement

- Ressources

- Lexique pondéré

- Conclusions / Perspectives

1 - Événements nominaux / EN Événement

Les événements nominaux

Les événements verbaux sont beaucoup traités en TAL

- Plus faciles à identifier
- Plus fréquents
- Mais aussi plus communs…

Les événements « importants » sont souvent nominalisés, mais :

- Ils ne portent pas toutes les informations véhiculées par le verbe
- Ils n'indiquent pas tous un événement !
- Même quand ils indiquent un événement, ils sont parfois ambigus

Un événement (chez nous) est « ce qui arrive », un changement d'état.

Les Entités Nommées

Étant donnés un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus.
[Ehrmann, 2008]

Dans Quaero :

ENTITÉS NOMMÉES

- Éléments « notables » des textes, comme les noms de personnes et de lieux

ENTITÉS NOMMÉES ÉTENDUES [Grouin *et al.*, 2011]

- Extension des entités nommées à de nouveaux types (e.g. civilisations, les fonctions, etc.)
- Extension de la définition des entités nommées à des expressions construites autour de noms communs : autorisation d'inclusion d'expressions ne contenant aucun nom propre

Les EN Événement

Les EN « classiques »

- monoréférentialité (Enjalbert, Vicente) ;
- entité du monde concret (Ester) ;
- définies comme des noms propres (MUC-7)

Le **festival de Cannes**

La **seconde guerre mondiale**

Les désignations nominales des événements

Le **festival** s'est déroulé en juillet.

L'**explosion d'un réacteur nucléaire**

Intérêt pour le TAL (cadre / perspectives de recherche)

Extraction d'information

Analyse d'un texte en surface dans le but d'une application précise

L'**explosion de la centrale** ≃ Le **festival de Cannes**

Exemple d'une application pratique : projet ANR Chronolines

Partenaires : MoDyCo, LIMSI, AFP, XRCE, Exalead

Buts :

- Ordonner les événements sur un axe temporel
- Proposer, dans la chronologie d'un thème, les événements les plus importants

Composition des événements nominaux

Trois types :

1. Des noms déverbaux / dérivés de verbes

   *Le 21 juin, c'est la **fête de la musique**.*
   *L'**adoption** par le Parlement d'une loi [...]*

2. Des noms qui évoquent des événements de façon ambiguë

   *L'**organisation du procès** dans les 60 jours*

   ou non ambiguë

   *le **Festival du film de Berlin***
   *le **Salon de l'Agriculture***

Composition des événements nominaux

3. Des mots qui prennent un caractère événementiel en contexte

(métonymie) ***Les frégates de Taïwan*** *s'invitent à Lorient.*

(toponyme) *Personne ne veut d'un nouveau* ***Tchernobyl****.*
Copenhague *se solde par un échec.*

(héméronyme) *Les indemnisations pour le* ***11 Septembre***
On pourrait assister à un ***21 avril*** *à l'envers.*

2 - Ressources

Corpus

Guide d'annotation EN Événement

Lexiques

Règles d'extraction

Analyseur syntaxique : XIP

Corpus

Corpus de type journalistique

Corpus manuellement annoté par les auteurs

Le Monde (2001-2002) + L'Est Républicain

	Le Monde	L'Est Républicain	total
articles	83	109	192
mots	31449	16197	47646
événements	**1105**	**736**	**1841**
événement/mots	3,51%	4,54%	3,86%

FR-TimeBank

L'Est Républicain 109 articles

Annotation : TimeML 663 événements nominaux

2 - Ressources

Corpus

Guide d'annotation EN Événement

Lexiques

Règles d'extraction

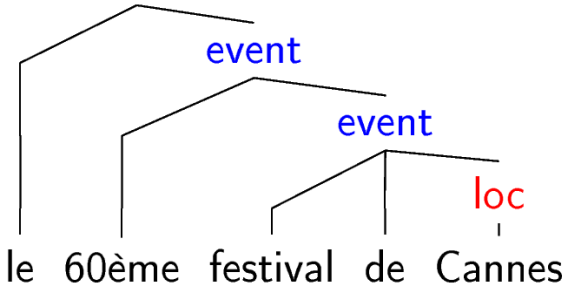
Analyseur syntaxique : XIP

Guide d'annotation

Complément au guide d'annotation en EN de Quaero

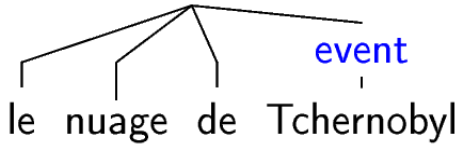
Annotations Imbriquées

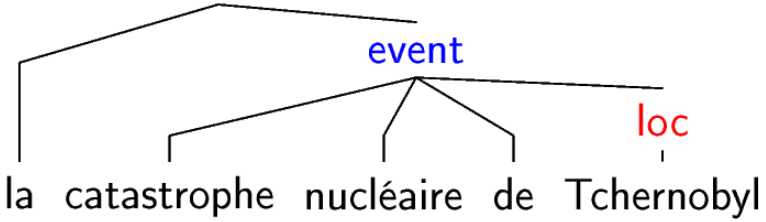
- Avec les autres entités
- Entre événements



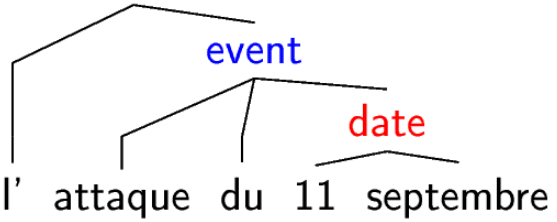
D'autres entités nommées devenues événement

- Lieux





- Date



Typologie 1/3

Modalité

Réel, réalisé → **type="factual"**

L'orientation d'une nouvelle expérience après <event type="factual" frequence="instance" temp="future"> ***les prochaines élections*** <event> *[...]*

Potentiel, hypothétique, probable → **type="hypothetical"**

pour ceux qui souhaitent une <event type="hypothetical" frequence="unique" temp="after"> ***victoire de la gauche*** </event>

Événement qui n'a pas eu lieu → **type="nonfactual"**

<event type="nonfactual" frequence="unique" temp="before"> ***Cette prétendue agression*** </event> *avait [...]*

Événement abstrait → **type="abstract"**

La <event type="abstract" frequence="unique" temp="now"> ***crise*** </event> *suit une période de confiance excessive.*

Typologie 2/3

Fréquence de l'événement

Unique → **frequence="unique"**

Et pour accroître le désarroi, on apprend qu' un <event type="hypothetical" frequence="unique" temp="before"> ***autre attentat*** </event> *se serait produit à moins de 100 km d'Alger.*

Récurrent → **frequence="recurring"**

Les <event type="abstract" frequence="recurring" temp="now"> ***Jeux paralympiques*** </event> *se tiennent toujours en marge de [...]*

Instanciation d'un phénomène récurrent → **frequence="instance"**

Les <event type="factual" freq="instance" temp="before"> <event type="abstract" type="recurring" temp="before"> ***Jeux Olympiques*** </event> ***de 1996*** </event> *ont été un succés.*

Typologie 3/3

Moment de la réalisation

Passé → **temp="before"**

Ils aboutissent à un <event type="factual" frequence="unique" temp="before"> **constat d'échec** </event>

Présent → **temp="now"**

<event type="factual" frequence="unique" temp="now"> **Cette rentrée-ci** </event> se place sous le signe de la contestation sociale.

Futur → **temp="after"**

Le <event type="factual" frequence="unique" temp="after"> **sommet Chine - Union Européenne** </event> se tiendra à Prague.

2 - Ressources

Corpus

Guide d'annotation EN Événement

Lexiques existants

Règles d'extraction

Analyseur syntaxique : XIP

Les lexiques existants

Lexique de noms déverbaux (nominalisations de verbes d'action ou de procès)

VerbAction [Tanguy et Hathout, 2002]

9393 couples verbes-lemme – 9200 lemmes nominaux uniques

(action/événement)	*ablation, sensibilisation, victimisation*
(ambigus)	*aération, étalage, shampooing, vœu*

Lexique complémentaire et agrémenté de noms ayant au moins une fois une interprétation événementielle

Lexique alternatif des noms événementiels [Bittar, 2010]

746 noms d'événement

(non-déverbaux)	*anniversaire, grêve*
(lexique spécifique)	*anticoagulothérapie*
(ambigus)	*apéritif*

Les lexiques existants

Deux difficultés :

Ambiguïté :

Les déverbaux peuvent désigner l'événement ou le résultat.

*La **construction** du port a duré 50 ans.* (événement)

*Cette **construction** fait 150 mètres de haut.* (résultat)

Difficulté même pour l'annotateur humain

*L'**étalage de marchandise** a été interdit sur le port.*

*Soumettre/ présenter une **proposition de loi***

Les mots autres que polysémiques qui prennent leur caractère événementiel en contexte (toponymes, héméronymes, etc.)

Constat : les lexiques seuls ne suffisent pas.

2 - Ressources

Corpus

Guide d'annotation EN Événement

Lexiques existants

Règles d'extraction

Analyseur syntaxique : XIP

Règles d'extraction

Les indicateurs temporels (IT)

Les événements sont ancrés dans le temps.

Ils peuvent être utilisés avec des prépositions temporelles, dans des compléments de temps.

le fait que l'événement se produise

à l'occasion de, lors de

A Jérusalem , lors de la ***réunion du gouvernement israélien*** *[...]*

usage référentiel de l'événement

pendant, la veille de, le lendemain de

*La population a été évacuée à la veille de l'****arrivée de la lave****.*

un moment de l'événement

à l'issue de, au commencement de

*les activistes qu'ils ont libérés au début de l'****Intifada****.*

Règles d'extraction

Les verbes d'événement et de cause/conséquence (VB)[Arnulphy *et al.*, 2010]

Des événements comme conséquence ou cause d'autres

Les ***crises*** ont pour origine des problèmes de défaillance technique.

Cette *élection* entraînera-t-elle la ***mise en sourdine*** **des intérêts communaux** ?

[...] *a provoqué* un ***tollé*** chez les organisations amérindiennes.

Les verbes qui introduisent des événements ou des noms d'événements

Le **Salon de l'Agriculture** est organisé Porte de Versailles

Les matches de huitièmes et de quart de finale ont eu lieu devant plus de 5000 personnes.

Le général, qui assistait à une ***cérémonie*** au côté du chef de l'État, s'est déclaré très content de sa décision.

2 - Ressources

Corpus

Guide d'annotation EN Événement

Lexiques existants

Règles d'extraction

Analyseur syntaxique : XIP

XIP

Analyseur syntaxique robuste XIP (XRCE)

[Aït-Mokhtar et al., 2002]

Dépendances syntaxiques

Reconnaissance d'EN (personne, lieu, organisation)

Possibilité pour l'utilisateur d'implémenter ses propres règles de grammaire en plus de celles existantes

Outil utilisé pour effectuer les expérimentations et évaluations

Évaluations des ressources

	Précision	Rappel	F-mesure
Évaluation des lexiques			
VerbAction	**48,7 %**	66,8 %	0,56 %
VerbAction + Bittar	**48,3 %**	84,1 %	0,61 %
Règles d'extraction (sans lexique)			
IT	81,2 %	**6,1 %**	0,11
VB90	**84,0 %**	1,1 %	0,02
VB90 + IT	**81,6 %**	**7,2 %**	0,13
Lexiques et règles			
Lexiques + Règles	**48 %**	85,9%	

3 - Lexique pondéré

Un lexique pondéré des noms d'événements

Rappel : Les règles d'extraction sont précises (> 80 %) mais très peu couvrantes (< 10 %).

Idée : lancer ces règles sur un gros corpus pour constituer un **lexique pondéré**

2 années du Monde (60 112 articles)

Pour chaque mot *m* extrait par les règles, on calcule le ratio entre

e(m) le nombre d'occurrences de ce mot extraites par les règles

t(m) le nombre d'occurrences totales de ce mot

$r(m) = e(m) / t(m)$

Pas une « probabilité », mais une valeur relative intéressante

Lexiques pondérés

Déclencheur des lexiques	Nb. détecté par les règles	Nb. total	Ratio
chute	434	2620	16,6 %
clôture	63	470	13,4 %
élection	1243	9713	12,8 %
guerre	1126	11542	9,8 %
crise	286	6185	4,6 %
expérience	63	2878	2,2 %
tension	16	1595	1,0 %
coopération	5	1631	0,3 %
subvention	2	867	0,2 %

Mots présents dans les lexiques disponibles

ambiguïté

Lexique pondéré

Déclencheur **(absent des lexiques)**	Nb. détecté par les règles	Nb. total	Ratio
Anschluss	3	4	75 %
méchoui	3	5	60 %
krach	20	169	11,8 %
RTT	14	166	8,4 %
demi-finale	35	553	6,3 %
cessez-le-feu	15	440	3,4 %
difficulté	16	3894	0,4 %
accès	9	2828	0,3 %
11 septembre	12	4354	0,3 %

ambiguïté

Mots absents des lexiques disponibles

Évaluation du lexique

Apprentissage : trois modèles basiques (classifieur à base de règles - J48) :

M_l : uniquement les deux lexiques « standards »

M_r : uniquement le lexique des ratios

M_{lr} : combinaison des lexiques « standards » et ratios

	M_l	M_r	M_{rl}
Precision	0.51	0.49	**0.54**
Recall	0.86	0.89	**0.89**
F-measure	0.64	0.63	**0.67**

3 – Conclusion / Perspectives

Conclusion

Évaluation des ressources :

- Les lexiques existants ne suffisent pas (15% d'événements manquants)

Un lexique :

- Constitué automatiquement
- Comportant une information supplémentaire : les ratios
- Conduit à des résultats identiques ou légèrement supérieurs aux autres lexiques (validés manuellement)
- Permet de récupérer des noms d'événements métonymiques (11 septembre, etc.) pour constituer une base d'événements « potentiels »

Perspectives

Lancer sur un corpus beaucoup plus gros pour une meilleure représentativité

L'utiliser dans un modèle plus abouti d'extraction des événements

Constituer un corpus d'apprentissage sur la base de mots non ambigüs choisis à la main (en contexte) :

« déclencheurs sûrs » « pas déclencheurs sûrs »

Passage à l'anglais