TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Un lexique pondéré des noms d'événements en français

Béatrice Arnulphy[1, 2] Xavier Tannier[1, 2] Anne Vilnat[1, 2]
(1) Univ. Paris-Sud 11, 91405 Orsay
(2) LIMSI-CNRS, 91403 Orsay
prenom.nom@limsi.fr

Résumé. Cet article décrit une étude sur l'annotation automatique des noms d'événements dans les textes en français. Plusieurs lexiques existants sont utilisés, ainsi que des règles syntaxiques d'extraction, et un lexique composé de façon automatique, permettant de fournir une valeur sur le niveau d'ambiguïté du mot en tant qu'événement. Cette nouvelle information permettrait d'aider à la désambiguïsation des noms d'événements en contexte[1].

Abstract. This article describes a study on automatic extraction of event nominals in French texts. Some existing lexicons are used, as well as some syntactic extraction rules, and a new, automatically built lexicon is presented. This lexicon gives a value concerning the level of ambiguity of each word as an event.

Mots-clés : extraction d'information, événements nominaux, lexiques.

Keywords: information extraction, nominal events, lexicons.

1 Introduction

La plupart des événements dans la langue est exprimée par les verbes et les noms. La forme verbale a été largement traitée, dans une approche formelle notamment par Vendler (1967) ou encore en traitement automatique des langues par le biais de TimeML (Pustejovsky *et al.*, 2005). Si les événements verbaux sont plus nombreux, plus simples à identifier et à lier aux autres informations temporelles, ils expriment souvent des événements plus communs et moins pertinents, tandis que la nominalisation d'un événement indique souvent son importance.

Les événements nominaux peuvent être construits de trois manières différentes. Certains sont construits à partir de noms déverbaux (*la **fête** de la musique* ou *l'**adoption** d'une réglementation*), sachant qu'ils peuvent désigner l'événement ou le résultat de l'action indiquée par le verbe dont il est issu (*construction*). D'autres sont formés à partir de noms autres que déverbaux et qui décrivent intrinsèquement des événements (*le **festival** de Cannes* ou *le **match** PSG-OM*), ces mots pouvant être ambigus (*le **salon** du livre*). Enfin, des syntagmes nominaux (SN) sans valeur événementielle peuvent par le résultat d'une métonymie, en contexte, référer à l'événement lié par exemple à un lieu (***Tchernobyl***), une date (*le **11 septembre***) ou l'objet d'une affaire (*les **frégates** de Taïwan*). Au vu de ces trois types d'événements, on se rend compte que le recours au lexique est nécessaire, mais pas suffisant ; une désambiguïsation par le contexte est indispensable.

L'événement est ce qui survient, le changement d'état opéré. Il peut être récurrent ou unique, prévu ou non, durer ou être instantané, se produire dans le passé, le présent ou le futur. Nous présentons une étude préalable détaillée des noms qui les caractérisent, par l'utilisation d'un corpus manuellement annoté. Après un bref état de l'art du domaine (section 2), nous présenterons les ressources dont nous disposons (section 3) et l'étude en vue de l'extraction des noms d'événements (section 4).

2 État de l'art

Quelques définitions des événements ont été proposées en philosophie, histoire, linguistique ou encore journalisme. Ces deux dernières disciplines nous intéressent tout particulièrement, parce que nous travaillons sur des

[1]Ce travail a été partiellement financé par OSEO dans le cadre du programme Quaero.

corpus d'articles et que les médias sont des inventeurs prolifiques de noms d'événements. Dans les années 70, à la suite de la théorie de Davidson concernant l'événement mental[2], ces travaux se focalisent sur « ce qui fait l'événement » et comment les **médias** le créent. Neveu & Quéré (1996) présentent l'événement comme une simple occurrence, non planifiée, qui ne se répète pas et qui se produit « dans un passé plus ou moins proche ». Rien n'y est dit au sujet des noms donnés à ces événements. En **linguistique**, par contre, quelques recherches ont essayé de traiter le problème des événements du point de vue des noms qu'on leur donne. Velde (2006) introduit par le parallèle entre les noms propres et la « triade je-ici-maintenant », la notion de nom propre de temps. De plus, les noms de lieux et les dates peuvent par le biais de la métonymie désigner des événements. C'est le cas pour le toponyme *Tchernobyl* (Lecolle, 2004) ou encore l'héméronyme *11 septembre* (Calabrese, 2008).

En **TAL**, la définition des événements semble fortement dépendante de l'application à laquelle elle est dédiée. TimeML (Pustejovsky *et al.*, 2003) est un formalisme d'annotation des événements principalement dédié aux événements verbaux. « Événement » est un « terme générique pour désigner les situations qui se produisent ». Dans nos travaux et par opposition à TimeML, nous n'acceptons pas les états et nous focalisons notre attention essentiellement sur les nominalisations. Par ailleurs, des campagnes d'évaluation ont intégré les événements. Automatic Content Evalution, ACE (Doddington *et al.*, 2004) a présenté un projet d'extraction, détaillé et précis,de catégories d'événements particuliers, liés à la vocation militaire du projet. Dans ESTER (Gravier *et al.*, 2004), on distingue les événements « historiques et uniques » et les « répétitifs », même si l'annotation des événements fut annulée. Plus récemment, en 2010, SemEval-2[3] s'est interessé aux événements.

L'**extraction automatique** des noms d'événement en français n'est pas encore étudiée. Pour d'autres langues, quelques recherches ont été vouées à cette thématique : Russo *et al.* (2011) s'est focalisée sur les moyens de repérer la lecture événementielle des noms déverbaux en italien et Creswell *et al.* (2006) se sont focalisés sur la désambiguisation des noms événementiels dans la version anglophone de WordNet. Dans la suite de ces travaux sur l'italien et l'anglais, nous proposons une étude pour l'extraction automatique des événements nominaux en français en nous aidant de ressources existantes et de celles élaborées pour les besoins de la tâche.

3 Description des ressources utilisées

3.1 Les corpus

Pour cette étude, deux corpus annotés ont été utilisés. Le premier, le FR-TimeBank, a été réalisé par Bittar (2010) dans le cadre de sa thèse, l'autre a été annoté par deux des auteurs même.

Le corpus FR-TimeBank est un ensemble de 109 articles du journal *L'Est Républicain* (ER), annoté selon le formalisme TimeML. Comme noté précédemment, le formalisme d'annotation TimeML ne s'arrête pas uniquement aux événements nominaux, mais ce corpus en compte tout de même 663.

Notre corpus manuellement annoté en événements nominaux Nous avons annoté manuellement 192 articles de presse des journaux *Le Monde* (LM – 83 articles) et *L'Est Républicain* (ER – les 109 articles utilisés dans le FR-TimeBank). Ce corpus représente 1844 noms d'événements[4]. Les annotations ont été conduites en fonction d'un guide d'annotation élaboré par nos soins. Ce document, qui n'est pas présenté ici, propose une typologie des événements ainsi que des instructions pour décider de l'événementialité d'un nom ou pas. Le but de ce corpus est d'annoter les SN utilisés pour nommer les événements. Pourtant dans le cadre de cette étude, nous ne nous concentrerons que sur les têtes de SN. En ce qui concerne les annotations de celles-ci, les deux annotateurs experts ont un bon accord (kappa = 0,808), sachant qu'environ 75 % du corpus a été annoté par les deux annotateurs.

Afin de comparer l'annotation Bittar et la nôtre, une sous-partie de notre corpus est le corpus FR-TimeBank. L'accord inter-annotateur calculé entre les deux annotations est bon : 0,704 (kappa). 79.8 % des têtes de SN d'événements sont étiquetées en commun dans nos deux corpus.

Enfin, à des fins d'expérimentation que nous détaillerons plus tard, nous avons eu recours à un corpus non-annoté composé de 60 112 articles de presse du *Monde* de l'année 2002, ne comptant aucun article du corpus annoté.

[2] Mental Events (1970), republié dans (Davidson, 1980)

[3] `http://semeval2.fbk.eu/`

[4] À titre de comparaison, le corpus italien IT-TimeBank (Russo *et al.*, 2011) compte 3695 événements, le FR-TimeBank 663 et le corpus (Creswell *et al.*, 2006) 1579.

3.2 Les lexiques

Deux lexiques ont été utilisés lors de nos expérimentations, il s'agit du VerbAction (Tanguy & Hathout, 2002) et du lexique alternatif des noms événementiels de Bittar (Bittar, 2010).

Le lexique **VerbAction** est constitué d'une liste de verbes d'action accompagnée des noms déverbaux morphologiquement apparentés à ceux-ci (9 393 couples verbe-nom, soit 9 200 lemmes nominaux uniques). Les verbes d'action impliquant que quelque chose se produise (*fêter*), les noms déverbaux de ces verbes devraient donc décrire une action (*fête*) et donc potentiellement nommer l'événement qui a lieu lorsque cette action se produit (*la fête de la musique*). Par ailleurs, ces mots peuvent donc être ambigus, principalement parce qu'ils servent aussi à référer au résultat des actions désignées par ces verbes (*aération*, *étalage*).

Le **lexique alternatif des noms événementiels** [5] se présente comme complémentaire au VerbAction dans la recherche de noms d'événements. Il contient 804 noms qui ne sont pas des déverbaux, comme *anniversaire* ou *grève*. Ces mots ont au moins une lecture événementielle dans le corpus étudié par leur auteur. Comme les déverbaux, certains mots sont ambigus : ils peuvent désigner l'événement / le procès aussi bien que le résultat ou l'objet de celui-ci, c'est le cas de *apéro* et *feu*. De plus, certains de ces noms présentent un état (*absence*), or les états ne font pas partie de notre définition de l'événement en tant que tel, nous envisageons l'événement sous le biais du changement d'état. Enfin, de nombreux noms appartiennent à des registres de langue spécifiques (*anticoagulothérapie*). Ce lexique a été utilisé comme indice pour l'annotation en TimeML du corpus FR-TimeBank.

3.3 Les règles d'extraction

La possibilité d'utiliser des **verbes d'événements et de cause/conséquence** pour l'extraction des noms d'événements a été étudiée par (Arnulphy *et al.*, 2010). Des verbes qui déclenchent des événements (en position sujet ou argument) ont été collectés. Nous reprenons les 15 verbes ayant une utilisation événementielle à plus de 90 % (VB90), en utilisant une approche syntaxique, contrairement à l'étude citée. Ces verbes sont : *avoir lieu*, *se produire*, *s'expliquer par*, *avoir pour origine*, *être entraîner* (en position sujet), et *organiser*, *déclencher*, *conduire à*, *assister à*, *donner lieu à*, *inciter à*, *occasionner*, *se précipiter à*, *tirer les conséquences de*, *tirer les leçons de* (en position d'argument).

De plus, les événements sont ancrés dans le temps et les noms donnés à ces événements peuvent être utilisés comme des entités temporelles et être par conséquent introduits par des **indicateurs temporels**. C'est pourquoi nous nous sommes intéressés à certains introducteurs temporels pour extraire des noms qui ne sont pas des dates et qui sont susceptibles d'indiquer des noms d'événements. Trois types de prépositions ont été utilisés. Elles indiquent le fait qu'un événement se produise : *à l'occasion de*, *lors de* (*À Jérusalem, **lors de la réunion** du gouvernement israélien*) ; l'usage référentiel de l'événement : *pendant*, *après*, *la veille de*, *le lendemain de* (*La population a été évacuée **avant l'arrivée** de la lave.*) ; encore un moment de l'événement : *à l'issue de*, *au commencement de* (*les activistes qu'ils ont libérés **au début de l'Intifada**.*) . Les règles d'extraction utilisant les indicateurs temporels sont nommées règles IT.

4 Expérimentations

Nous avons utilisé l'analyseur syntaxique XIP pour mener à bien nos expérimentations sur les lexiques et règles d'extraction verbes et indicateurs temporels.

XIP (Aït-Mokhtar *et al.*, 2002) est un analyseur syntaxique robuste qui permet une analyse pour le français et l'anglais des relations de dépendances et de la reconnaissance d'entités nommées classiques. Les événements n'y sont pas traités. Ce produit développé par Xerox Research Centre Europe est distribué avec des grammaires encryptées, inaccessibles à l'utilisateur. Pourtant, il est possible d'enrichir l'analyse par l'ajout de ressources et la création de ses propres règles de grammaire.

[5] Nous remercions André Bittar d'avoir mis à notre disposition son lexique complémentaire de noms d'événements.

4.1 Évaluation des lexiques

Les deux lexiques décrits en section 3.2 sont utilisés pour annoter les textes. Nous ne marquons comme événements que les mots du lexique qui ont le même lemme et qui sont de nature nominale. Les résultats sur notre corpus annoté sont présentés dans le tableau 1.

	Précision	Rappel	F-mesure
LM	43,7 %	65,1 %	0,52
ER	58,3 %	69,1 %	0,63
LM + ER	**48,7 %**	**66,8 %**	**0,56**

a) VerbAction

	Précision	Rappel	F-mesure
LM	43,9 %	84,3 %	0,58
ER	57,0 %	84,0 %	0,68
LM + ER	**48,3 %**	**84,1 %**	**0,61**

b) VerbAction + Bittar

TAB. 1 – Résultats avec les lexiques VerbAction et Bittar

Nous observons tout d'abord que pour les différents corpus (LM ou ER) les résultats en termes de précision et de rappel sont homogènes. VerbAction, appliqué seul, obtient 48,7 % de précision sur notre corpus entier ce qui nous montre que les déverbaux ont plus souvent une lecture non-événementielle. Le rappel s'élevant pour ce lexique à 66,8 %, nous pouvons conclure qu'environ un tiers des noms d'événements de notre corpus ne sont pas construits à partir de déverbaux. Lors de la combinaison du VerbAction avec le lexique alternatif de Bittar, le rappel est augmenté (de 66,8 % à 84,1 %) sans affecter la précision (de 48,7 % à 48,3 %). Mais 15 % des noms d'événements sont toujours absents et la précision est plutôt basse.

4.2 Utilisation des règles d'extraction (sans lexique)

Le tableau 2 donne les résultats de l'expérimentation sur les règles présentées dans la section 3.3, implémentées au moyen de XIP.

	Précision	Rappel	F-mesure
IT	81,2 %	**6,1 %**	0,11
VB90	**84,0 %**	1,1 %	0,02
VB90 + IT	**81,6 %**	**7,2 %**	0,13
VB75	68,2 %	3,2 %	0,06
VB75 + IT	76,2 %	9,2 %	0,16

TAB. 2 – Résultats avec les règles d'extraction IT et VB sur notre corpus annoté

Étant donné que les règles sont extrêmement restrictives et les verbes utilisés des déclencheurs sûrs à 90%, il n'est pas étonnant que contrairement à l'approche motivée par les lexiques, les règles d'extraction ont une bonne précision (toujours supérieure à 80 %) et un mauvais rappel (entre 1 % pour les règles à base de VB90 suivant leur configuration événementielle préférée et 7,2 % pour la combinaison IT et VB90). À titre de comparaison, nous avons mené la même expérience en utilisant les verbes qui ont présenté des groupes nominaux événementiels à plus de 75 % (dans l'étude précédemment citée). Nous pouvons nous rendre compte que les résultats ne sont pas meilleurs : un rappel de 2 points plus élevé pour une précision de 16 points de moins pour les verbes seuls et 5 de moins pour la combinaison avec les règles IT.

4.3 Combinaison des règles et des lexiques

Comme nous le voyons, l'utilisation des lexiques et des règles est *a priori* destinée à deux buts distincts. Les règles, appliquées sur le corpus de test, n'ont pas vocation à extraire de nouveaux mots qui seraient absents des lexiques, mais plutôt à confirmer qu'un mot appartenant à un lexique est bien un événement. Ceci est confirmé par le fait qu'ajouter les règles aux lexiques dans la phase de test n'améliore que très peu le rappel : de 48,3 % de précision à 48 % et de 84,1 % de rappel à 85,9 %. Cependant, comme nous allons le voir à la section suivante, ces règles peuvent permettre la constitution d'un lexique enrichi et pondéré, lorsqu'elles sont utilisées sur un nombre élevé de documents.

4.4 Un lexique pondéré des noms d'événements potentiels

L'évaluation des règles implémentées avec XIP étant tout à fait satisfaisante (précision supérieure à 80 %), celles-ci peuvent servir à constituer un lexique de façon automatique, à condition d'être appliquées sur un grand corpus, en l'occurrence un an du *Monde* (présenté à la section 3.1). Qui plus est, ce nouveau lexique contient également une information concernant le degré d'ambiguïté (événement *vs* non-événement) de chaque mot pour le corpus.

Après avoir appliqué les règles sur le corpus, nous notons, pour chaque nom, le nombre d'occurrences étiquetées par les règles, et le nombre d'occurrences total dans les documents. Nous obtenons un ratio qui, s'il ne représente bien sûr pas une proportion ou une probabilité de l'usage événementiel du mot (les règles conduisant à un faible rappel), permet, en comparaison des ratios de l'ensemble des mots, d'estimer son degré d'ambiguïté.

Ceci est illustré par les exemples donnés dans le tableau 3. Beaucoup de ces mots sont présents dans les lexiques *VerbAction* ou *Bittar* (colonne de gauche), d'autres en sont absents (colonne de droite). On voit que les noms peu ou pas du tout ambigus (ceux déclenchant toujours un événement) ont un ratio relativement élevé (supérieur au rappel moyen décrit dans la section précédente). C'est le cas de *chute*, *élection* ou *krach*. On note que *clôture*, fortement ambigu dans le cas général, l'est semble-t-il beaucoup moins dans des articles de journaux, où l'objet est bien moins évoqué que le fait de clore. En revanche, des mots comme *tension*, *subvention* ou *accès* sont très ambigus et obtiennent un ratio peu élevé. C'est également le cas de la date *11 septembre*, mais celle-ci est une des seules dates du lexique obtenu, et a de loin le meilleur ratio.

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 %

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 %

TAB. 3 – Exemples de déclencheurs potentiels collectés par les règles d'extraction.

Dans une dernière expérimentation, nous avons appliqué ce nouveau lexique obtenu automatiquement, pour le comparer aux lexiques VerbAction et Bittar. Nous avons utilisé différentes « tranches » de ce lexique, selon les ratios obtenus, pour observer l'évolution des performances : tous les mots ayant un ratio supérieur à 10 %, puis tous ceux ayant un ratio supérieur à 8 %, 6%, etc. Les résultats sont présentés dans le tableau 4.

Mots dont le ratio est supérieur à	Précision	Rappel	F-mesure
10 %	84,1 %	16,6 %	0,28
8 %	83,6 %	24,3 %	0,38
6 %	79,8 %	31,5 %	0,45
1 %	56,3 %	71,0 %	0,63
0,5 %	43,4 %	80,1 %	0,56

TAB. 4 – Application du lexique automatique par « tranches » de ratio.

La précision et le rappel évoluent bien entendu de manière opposée (lorsque le lexique est moins sélectif, le rappel augmente et la précision diminue), et la meilleure F-mesure (pour 1 %) est de 0,63, soit une valeur similaire à la F-mesure des deux lexiques VerbAction et Bittar combinés (0,61). Nous obtenons donc, automatiquement, un lexique de qualité comparable aux deux lexiques composés de façon semi-automatique, en ajoutant en plus l'information du degré d'ambiguïté sur la lecture événementielle des noms.

5 Conclusion

Dans cet article, nous avons présenté plusieurs expérimentations dans le cadre d'une étude sur les événements nominaux dans les textes en français. Nous avons notamment utilisé les lexiques disponibles et créé un nouveau lexique d'événements « potentiels », qui associe à chaque nom une valeur indiquant à quel point ce mot peut avoir une lecture événementielle. Ceci devrait aider à la phase de désambiguïsation en contexte, qui est indispensable pour une extraction efficace des événements nominaux. Cette phase de désambiguïsation est notre prochaine étape de travail, grâce à l'ensemble des lexiques et des indices contextuels collectés dans nos différentes études.

Références

AÏT-MOKHTAR S., CHANOD J.-P. & ROUX C. (2002). Robustness beyond Shallowness : Incremental Deep Parsing. *Natural Language Engineering*, **8**, 121–144.

ARNULPHY B., TANNIER X. & VILNAT A. (2010). Les entités nommées événement et les verbes de cause-conséquence. In *Actes de TALN 2010*, Montreal, Canada.

BITTAR A. (2010). *Construction d'un TimeBank du français : Un corpus de référence annoté selon la norme ISO-TimeML*. PhD thesis, Université Paris Diderot - École doctorale de Sciences du Langage / Laboratoire ALPAGE.

CALABRESE L. (2008). Les héméronymes. Ces évènements qui font date, ces dates qui deviennent évènements. *Mots. Les langages du politique*, **3**, 115–128.

CRESWELL C., BEAL M. J., CHEN J., CORNELL T. L., NILSSON L. & SRIHARI R. K. (2006). Automatically extracting nominal mentions of events with a bootstrapped probabilistic classifier. In *Proceedings of the COLING/ACL on Main conference poster sessions*, COLING-ACL '06, p. 168–175, Sydney, Australia : Association for Computational Linguistics.

DAVIDSON D. (1980). *Essays on Actions and Events*, chapter 11 "Mental Events" (1970). Psychology as Philosophy. Calendron Press : Oxford, UK.

DODDINGTON G., MITCHELL A., PRZYBOCKI M., RAMSHAW L., STRASSEL S. & WEISCHEDEL R. (2004). The Automatic Content Extraction (ACE) Program - Tasks, Data, and Evaluation. In *Proceedings of LREC'04*, Lisbonne, Portugal.

GRAVIER G., BONASTRE J.-F., GEOFFROIS E., GALLIANO S., MCTAIT K. & CHOUKRI K. (2004). ESTER, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en français. In *Proceedings of JEP'04*, Fèz, Maroc.

LECOLLE M. (2004). Toponymes en jeu : Diversité et mixage des emplois métonymiques de toponymes. In *Studii si cercetari filologice 3 / 2004, Université de Pitesti, Roumanie*.

NEVEU E. & QUÉRÉ L. (1996). Présentation. *Réseaux*, **14**(75), 7–21.

PUSTEJOVSKY J., CASTAÑO J., INGRIA R., SAURÍ R., GAIZAUSKAS R., SETZER A. & KATZ G. (2003). TimeML : Robust Specification of Event and Temporal Expressions in Text. In *IWCS-5, Fifth International Workshop on Computational Semantics.*

PUSTEJOVSKY J., KNIPPEN R., LITTMAN J. & SAURÍ R. (2005). Temporal and Event Information in Natural Language Text. *Language Resources and Evaluation*, **39**(2-3), 123–164.

RUSSO I., CASELLI T. & RUBINO F. (2011). Recognizing deverbal events in context. In *Proceedings of 12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2011), poster session*, Tokyo, Japan : Springer.

TANGUY L. & HATHOUT N. (2002). Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web. In J.-M. PIERREL, Ed., *Actes de TALN 2002*, p. 245–254, Nancy : ATALA ATILF.

VELDE D. V. D. (2006). *Grammaire des événements*. Sens et structures. Presses Universitaires du Septentrion.

VENDLER Z. (1967). *Facts and events*, chapter Verbs and Times, p. 97–121. Cornell University Press : Ithaca, NY, USA.