Combinaison d'approches pour l'extraction automatique d'événements

Laurie Serrano, Thierry Charnois, Stephan Brunessaux, Bruno Grilheres, Maroua Bouzid

Combinaison d'approches pour l'extraction automatique d'événements

Laurie Serrano[1, 2] Thierry Charnois[1] Stephan Brunessaux[2]
Bruno Grilheres[2] Maroua Bouzid[1]
(1) Laboratoire GREYC, Université de Caen Basse-Normandie
Campus Côte de Nacre, Boulevard du Maréchal Juin, BP 5186 - 14032 Caen
(2) Département IPCC, Cassidian
Parc d'Affaires des Portes - 27600 Val de Reuil
prenom.nom@unicaen.fr, prenom.nom@cassidian.com

RÉSUMÉ

Dans cet article, nous présentons un système d'extraction automatique d'événements fondé sur deux approches actuelles en extraction d'information : la première s'appuie sur des règles linguistiques construites manuellement et la seconde se fonde sur un apprentissage automatique de patrons linguistiques. Les expérimentations réalisées montrent que combiner ces deux méthodes d'extraction permet d'améliorer significativement la qualité des événements extraits (amélioration de près de 10 points de F-mesure).

ABSTRACT

Automatic events extraction by combining multiple approaches

In this paper, we present an automatic system for extracting events based on the combination of two existing information extraction approaches : the first one is made of hand-crafted linguistic rules and the second one is based on an automatic learning of linguistic patterns. We have shown that this mixed approach leads to a significant improvement of extraction performances.

MOTS-CLÉS : Extraction d'information, événements, approche symbolique, apprentissage de patrons linguistiques.

KEYWORDS: Text mining, events, symbolic extraction, linguistic pattern learning.

1 Introduction

Face à l'augmentation vertigineuse des informations disponibles librement (en particulier sur le Web), repérer efficacement celles qui peuvent nous intéresser s'avère une tâche longue et complexe. En réponse à cela, l'équipe IPCC [1] développe le WebLab [2], une plateforme d'intégration de services de "media mining" [3] pour la découverte de connaissances et l'aide à la décision. Nous présentons dans cet article une étude comparative de trois approches d'extraction d'événements : une approche symbolique basée sur des règles linguistiques construites manuellement, une méthode d'apprentissage de patrons linguistiques et une approche mixte. Les expériences menées

1. Information Processing, Control and Cognition, Cassidian
2. http://weblab-project.org/, consulté le 21/03/2012
3. Fouille de documents multimédia

montrent que la combinaison des deux premières méthodes permet d'améliorer significativement la qualité des événements extraits. Les travaux présentés ici sont réalisés dans le cadre de l'élaboration d'un système plus global de capitalisation des connaissances exploitant les technologies du Web sémantique (Serrano *et al.*, 2012). Précisons également que les événements que nous souhaitons extraire ont été au préalable définis dans une ontologie de domaine (nommée WOOKIE [4]). Nous proposons, dans un premier temps, un rapide tour d'horizon des travaux existants en extraction d'événements. Puis, nous décrivons les trois approches proposées, les expérimentations mises en place, nos premiers résultats et les perspectives envisagées.

2 Extraction des événements : tour d'horizon

L'extraction d'information est une discipline récente qui consiste à analyser un texte de manière automatique afin d'en extraire un ensemble d'informations jugées pertinentes (Poibeau, 2003). Deux approches dites "classiques" émergent : l'extraction basée sur des techniques linguistiques et les systèmes statistiques. Les tâches les plus communes en extraction d'information sont l'extraction d'entités nommées (Nadeau et Sekine, 2007), de relations entre entités et d'événements. L'extraction d'événements est particulièrement utilisée dans les activités de veille économique et stratégique ((Capet *et al.*, 2011) pour la détection de crise). Celle-ci peut-être conçue comme une forme particulière d'extraction de relations où une "action" est liée à d'autres entités telles qu'une date, un lieu, des participants, *etc*. Plusieurs campagnes MUC [5] s'y sont intéressé avec notamment des tâches de remplissage automatique de formulaires ("template filling"). Comme en extraction d'information de façon générale, la littérature du domaine offre à la fois des travaux basés sur des approches symboliques et des techniques purement statistiques. (Aone et Ramos-Santacruz, 2000) développe REES, un extracteur d'événements basé sur des règles linguistiques construites manuellement couplées à une analyse syntagmatique. Dans la lignée, (Grishman *et al.*, 2002) s'intéresse à la détection d'événements épidémiques au moyen d'un transducteur à états finis. Toutefois, ces méthodes purement linguistiques, bien que généralement très précises, ont pour principales faiblesses d'être spécifiques à un domaine donné, d'avoir un taux de rappel plutôt faible et un coût de développement manuel élevé. Du côté des approches statistiques, (Ahn, 2006) propose de combiner plusieurs classifieurs pour l'extraction des événements dans la campagne ACE. L'apprentissage statistique permet de prévoir de nombreux contextes d'apparition mais nécessite une grande quantité de données annotées pour être performant et construit un modèle de type "boite noire" non-accessible et non-modifiable. Face à cela, les méthodes d'apprentissage de patrons ou les approches semi-supervisées apparaissent intéressantes comme par exemple le système de (Xu *et al.*, 2006). Observant que toutes ces approches prises séparément restent imparfaites, nous proposons d'élaborer une approche hybride permettant d'exploiter les points forts des méthodes "classiques". Pour cela, nous avons choisi, de compléter les performances d'un extracteur d'événements symbolique par un système d'apprentissage de patrons linguistiques.

3 Modélisation des événements

L'événement étant l'objet central de nos travaux, il est nécessaire de définir plus précisément ce concept. Considéré comme une entité aux propriétés spécifiques, l'événement a particulièrement

4. Weblab Ontology for Open sources Knowledge and Intelligence Exploitation
5. Message Understanding Conference

été étudié en philosophie (Davidson, 2001) et en linguistique (Van De Velde, 2006). Après avoir considéré différents travaux, nous prenons pour point de départ la définition de (Krieg-Planque, 2009) qui nous paraît adaptée : "un événement est une occurrence perçue comme signifiante dans un certain cadre". Afin de proposer une représentation plus formelle d'un événement, nous nous appuyons sur les travaux de (Saval *et al.*, 2009) qui propose une extension sémantique pour la modélisation des événements de type "catastrophes naturelles". Celui-ci définit un événement E comme la combinaison de 3 composantes : une propriété sémantique S, un intervalle temporel I, et une entité spatiale SP. Un événement est donc représenté sous la forme $E\langle I, SP, S\rangle$. Dans notre cas, la propriété sémantique est définie par les différents types d'événement de notre ontologie (que nous décrivons plus loin), la composante temporelle constitue la date ou période d'occurrence d'un événement et l'entité spatiale correspond à son lieu d'occurrence. Nous proposons d'adapter cette représentation à notre domaine d'application en l'enrichissant d'une composante supplémentaire A correspondant aux différents participants impliqués dans l'événement. Nous avons donc maintenant $E\langle I, SP, S, A\rangle$ où A est un ensemble de participants jouant un ou plusieurs rôle(s). Un participant est noté P_i où $0 \leq i < n$ et un rôle est noté r_j où $0 \leq j < k$. La composante A est donc définie de la façon suivante : $A = \{(P_\alpha, r_\beta)\}$ tel que le participant P_α joue le rôle r_β dans l'événement en question. Cette modélisation a été implémentée au sein de notre ontologie de domaine WOOKIE, centrée sur 5 classes supérieures : "Event", "Person", "Unit", "Place" et "Equipment". Les différentes approches comparées dans cet article visent à extraire la vingtaine d'événements suivants : "AttackEvent", "BombingEvent", "ShootingEvent", "CrashEvent", "DamageEvent", "DeathEvent", "FightingEvent", "InjureEvent", "KidnappingEvent", "MilitaryOperation", "ArrestOperation", "HelpOperation", "PeaceKeepingOperation", "SearchOperation", "SurveillanceOperation", "TrainingOperation", "TroopMovementOperation", "NuclearEvent", "TrafficEvent".

4 Extraction des événements : approches proposées

4.1 Approche à base de règles linguistiques

L'approche que nous présentons ici a été implémentée grâce à la plateforme d'ingénierie textuelle GATE [6] et vise à extraire un ensemble d'événements tels que définis ci-dessus ainsi que les participants et circonstants suivants : la date de l'événement, son lieu d'occurrence et les entités de type "personne" et "organisation" impliquées. La figure 1 résume les différentes étapes de notre approche (le repérage des rôles ne sera pas traité dans cet article, pour une description détaillée se référer à (Serrano *et al.*, 2011)). Notre système repose sur une chaine de traitement composée de différents modules d'analyse linguistique ("tokenisation", découpage en phrases, repérage lexical, étiquetage grammatical, analyse syntaxique, transducteur à états finis, *etc*.). Nous définissons tout d'abord un ensemble de termes considérés comme possibles déclencheurs d'événement (dits "noms d'événement"). Nous choisissons de nous limiter, pour l'instant, aux déclencheurs verbaux et nominaux et de constituer des listes de lemmes, plus courtes et permettant d'étendre le repérage à toutes les formes fléchies. Ces déclencheurs (139 lemmes actuellement) sont répartis en différentes listes, chacune étant associée à un type d'événement (c'est-à-dire à une classe d'événement de notre ontologie) afin d'être repérés et annotés dans le corpus à analyser. Nous associons ensuite à ces "noms d'événement" les différentes entités qu'ils impliquent. Pour

6. General Architecture for Text Engineering

cela, nous effectuons, dans un premier temps, une extraction automatique d'entités nommées [7] ainsi qu'une analyse en constituants syntaxiques (syntagmes verbaux, nominaux, *etc.*). Enfin, ces différents éléments sont rattachés au "nom d'événement" grâce à une analyse syntaxique en dépendance (réalisée par le Stanford Parser [8]) ainsi que des règles de grammaire élaborées manuellement (dans le langage JAPE [9]). A l'heure actuelle, nous obtenons une annotation positionnée sur le "nom d'événement" résumant ses différents participants et circonstants et indiquant pour chacun d'eux s'il correspond à une entité nommée détectée précédemment.

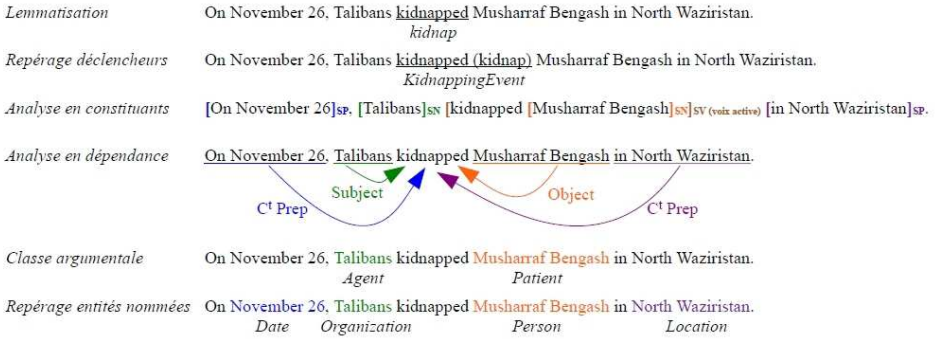

FIGURE 1 – Chaine d'extraction d'événements pour l'anglais

4.2 Apprentissage de patrons linguistiques

Par ailleurs, nous nous intéressons à l'extraction d'événements par une technique d'extraction de motifs séquentiels fréquents. Ce type d'approche permet d'apprendre automatiquement des patrons linguistiques compréhensibles et modifiables par un expert linguiste. La découverte de motifs séquentiels a été introduite par (Agrawal *et al.*, 1993) dans le domaine du "data mining" et adaptée par (Cellier et Charnois, 2010) à l'extraction d'information dans les textes. Ceux-ci s'intéressent en particulier à l'extraction de motifs séquentiels d'itemsets. Il s'agit de repérer, dans un ensemble de séquences, des enchainements d'items ayant une fréquence d'apparition supérieure à un seuil donné (dit "support"). La recherche de ces motifs s'effectue dans une base de séquences ordonnées d'itemsets où chaque séquence correspond à une unité de texte (ici la phrase). Un itemset est un ensemble d'items décrivant un mot de cette séquence. Un item correspond à une caractéristique particulière de ce mot telle que la catégorie grammaticale, le lemme, la forme fléchie, *etc*. Un certain nombre de paramètres peuvent être adaptés selon l'application visée : nature de la séquence et des items, nombre d'items, support, *etc*. La fouille sur un ensemble de séquences d'itemsets permet l'extraction de motifs combinant plusieurs types d'item et d'obtenir ainsi des patrons génériques, spécifiques ou mixant les informations (ce qui n'est pas permis par les motifs d'items simples), comme par exemple les patrons suivants : <homme de culture> <homme de N> <N PRP N> [10], *etc*. De plus, contrairement aux différentes approches que nous venons de mentionner, l'apprentissage de patrons ne nécessite ni corpus annoté avec les entités-cibles, ni analyse syntaxique. Cela constitue un réel avantage car, tout

7. Extraction des dates, lieux, personnes et organisations réalisée par une chaine GATE (Serrano *et al.*, 2011)
8. `http://nlp.stanford.edu/software/lex-parser.shtml`
9. Java Annotation Patterns Engine
10. N pour la catégorie nom, PRP pour préposition

d'abord, l'annotation manuelle de corpus reste un effort important et l'analyse syntaxique est encore une technologie aux performances inégales et peu disponible librement selon les langues. Le point faible partagé par toutes ces méthodes d'apprentissage symbolique reste le nombre important de motifs extraits. Pour pallier ce problème, (Cellier et Charnois, 2010) propose l'ajout de contraintes pour diminuer la quantité de motifs retournés. Dans la lignée de ces travaux, nous utilisons l'outil d'extraction de motifs séquentiels développé au GREYC (Béchet *et al.*, 2012). Celui-ci présente plusieurs points forts : il extrait uniquement des motifs dits "clos" (c'est-à-dire non redondants) et génère ainsi moins de motifs que d'autres systèmes. De plus, ce logiciel s'avère robuste et permet la fouille de séquences d'itemsets, fonctionnalité qui est rarement proposée par les outils existants. Nous avons adapté la fouille de motifs à notre domaine d'application et au traitement de dépêches de presse dans le but d'obtenir des patrons linguistiques permettant la détection d'événements. Ainsi, notre approche propose tout d'abord de pré-traiter un corpus grâce à l'outil TreeTagger [11] afin d'obtenir un découpage en séquences (ici en phrases) ainsi que différents types d'items : forme fléchie, lemme, catégorie grammaticale. Elle nécessite également une annotation sémantique en entités nommées. Enfin, nous effectuons un repérage lexical des "noms d'événement" et de leur type. Comme prévu, le nombre de motifs retournés par l'outil s'avère élevé, nous introduisons donc un ensemble de contraintes spécifiques à notre application : des contraintes linguistiques d'appartenance (nous pouvons par exemple choisir de ne retourner que des motifs contenant au moins un "nom d'événement" et une date) mais aussi une contrainte dite de "gap" (Dong et Pei, 2007), autorisant l'extraction de motifs ne contenant pas nécessairement des itemsets consécutifs (contrairement aux n-grammes dont les éléments sont strictement contigus). Ainsi un "gap" d'une valeur maximale n signifie qu'au maximum n itemsets (mots) sont présents entre chaque itemset du motif dans les séquences correspondantes. Cette approche non-supervisée nécessite une sélection manuelle des motifs pertinents. Pour cela, nous utilisons l'outil Camelis (Ferré, 2009) permettant d'ordonner et visualiser les motifs des plus généraux aux plus spécifiques puis de filtrer les plus pertinents. Les motifs ainsi sélectionnés sont ensuite appliqués sur un nouveau corpus afin d'en extraire les relations visées.

4.3 Vers une approche mixte

Nous travaillons actuellement à définir une méthode d'hybridation qui permette d'exploiter les forces de chacune des approches présentées. En effet, comme nous l'avons déjà souligné, l'extraction d'information à base de règles écrites manuellement s'avère généralement très précise mais peu couvrante alors que les techniques d'apprentissage montrent habituellement un meilleur rappel. Nous proposons donc, dans un premier temps, d'effectuer une simple union des résultats des deux extracteurs afin de maximiser le rappel de notre approche mixte (les expérimentations reportées dans cet article sont basées sur cette approche).

5 Expérimentations

Les expérimentations présentées consistent à extraire un ensemble d'événements d'un corpus journalistique en anglais et dont le thème est d'intérêt pour le renseignement. Nous nous focalisons sur une vingtaine de types d'événement (*cf* section 3) et sur les relations suivantes : la date de l'événement, son lieu d'occurrence et les actants impliqués (personnes et organisations).

11. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

5.1 Corpus et paramètres d'apprentissage

Cette évaluation a nécessité les deux corpus suivants :

– Le premier corpus est un corpus d'apprentissage nécessaire à la mise en place de l'outil d'extraction de motifs séquentiels fréquents : il s'agit d'un corpus de textes anglais abordant une thématique militaire et annoté avec des entités nommées et des "noms d'événement". Nous avons constitué ce corpus de manière semi-automatique à partir de 400 dépêches de presse sur l'engagement du Canada en Afghanistan [12] et de 700 dépêches parues entre 2003 et 2009 sur le site de l'ISAF [13]. Ce corpus a, dans un premier temps, été annoté automatiquement en entités nommées et "noms d'événement" grâce à notre outil d'extraction basé sur GATE (Serrano *et al.*, 2011), puis nous avons revu manuellement ces annotations pour corriger les éventuels erreurs/oublis et ainsi garantir la qualité des données d'apprentissage.
– Le second corpus est un corpus de test permettant de comparer notre extraction d'événements par rapport à une vérité-terrain. Pour cela nous avons choisi d'utiliser un corpus fourni dans la campagne d'évaluation MUC-4 et constitué de 100 dépêches de presse relatant des faits terroristes en Amérique du Sud. Notre évaluation porte sur une partie de ce corpus annotée manuellement [14], soit environ 210 événements et près de 240 relations (55 relations de type "date", 65 relations de type "lieu" et 120 relations de type "participant").

Pour mettre en place notre approche d'apprentissage symbolique, nous avons, tout d'abord, opéré un apprentissage de motifs séquentiels fréquents sur le premier corpus en considérant quatre caractéristiques (quatre types d'item) : la forme fléchie du mot, sa catégorie grammaticale, son lemme et sa classe sémantique ("nom d'événement", "date", "lieu", "personne" ou "organisation"). Nous avons choisi de réaliser une tâche d'apprentissage par type d'entité impliquée en utilisant le système des contraintes d'appartenance proposé par l'outil de (Béchet *et al.*, 2012). Nous obtenons donc quatre séries de motifs de type "nom d'événement"-"date", "nom d'événement"-"lieu", "nom d'événement"-"personne" et "nom d'événement"-"organisation". Nous avons également procédé à plusieurs essais de paramétrage et, au regard de ces tests, nous avons choisi de fixer un "gap" maximal de 3 itemsets (correspondant à 3 mots possibles entre chaque élément du motif) et un support absolu relativement bas (10 en valeur absolue, soit 6% des séquences pour tous les types de relation) afin d'obtenir des motifs intéressants mais en nombre raisonnable pour une exploration et une validation manuelles (environ 12000 motifs au total).

5.2 Résultats et discussion

Nous avons réalisé manuellement une comparaison des extractions obtenues par chacune des deux approches (appliquée séparément) et celles résultant de l'approche mixte. Le tableau 1 présente les scores de précision, rappel et F-mesure de chaque approche, globalement et par type de relation. Précisons que ces résultats proviennent d'une extraction de relations fondée sur l'annotation manuelle des entités nommées que nous avons réalisée sur le corpus de test (et non pas sur une extraction automatique) afin d'éviter que des erreurs dans l'extraction des entités viennent perturber l'extraction de relations. Nous pouvons constater que l'approche à base de règles et l'apprentissage de motifs obtiennent tous deux une très bonne précision globale et que,

12. `http://www.afghanistan.gc.ca/canada-afghanistan`, consulté le 21/03/2012
13. `http://www.nato.int/isaf/docu/pressreleases`, consulté le 21/03/2012
14. Nous avons choisi de ne pas réutiliser les "templates" de référence fournis avec le corpus car le nombre et le type des événements ne correspondaient pas à notre modélisation et ne permettaient pas d'évaluer la totalité de nos extractions.

	Approche à base de règles manuelles			Apprentissage de motifs			Approche mixte		
	Précision	*Rappel*	*F-mesure*	*Précision*	*Rappel*	*F-mesure*	*Précision*	*Rappel*	*F-mesure*
Date	0,93	0,25	0,39	0,90	0,64	0,75	0,90	0,68	0,78
Lieu	0,92	0,37	0,53	0,86	0,49	0,63	0,81	0,60	0,69
Participants	0,97	0,49	0,42	0,93	0,32	0,47	0,92	0,51	0,66
Toutes relations	0,94	0,37	**0,45**	0,90	0,48	**0,62**	0,88	0,60	**0,71**

TABLE 1 – Extraction d'événements : évaluation des trois approches

comme attendu, le rappel est meilleur pour cette dernière approche. Par ailleurs, nous avons été assez surpris par la bonne précision de la méthode par apprentissage, que nous expliquons par une sélection manuelle restrictive et précise des motifs. Les taux de rappel obtenus sont peu élevés : ce résultat est conforme à l'état de l'art pour l'approche à base de règles. Pour l'approche à base d'apprentissage, l'utilisation d'un "gap" maximal trop restreint ne permet pas d'extraire les relations distantes. Ce qu'il faut retenir de ces expérimentations est que l'approche mixte obtient une F-mesure nettement supérieure (près de 10 points par rapport à la meilleure des deux approches), ce qui dénote une amélioration globale de la qualité d'extraction pour tout type de relation. De plus, nous remarquons que l'apprentissage de patrons complète avec succès notre approche symbolique en augmentant sensiblement le taux global de rappel. Nous constatons cependant une légère perte de précision qui résulte du nombre plus élevé de règles et patrons linguistiques au sein de l'approche mixte entraînant une augmentation des faux positifs.

Apport de l'analyse syntaxique Parallèlement à ces résultats, nous nous sommes intéressés à l'apport de l'analyse syntaxique au sein de notre approche mixte : les résultats du tableau sont issus de notre système avec analyse en dépendance, sans cela nous aurions eu une perte de performances considérable (11 points de F-mesure, 19 points de précision et 1 point de rappel). Bien que les outils d'analyse syntaxique soient inégalement disponibles selon les langues, cette observation confirme l'intérêt de cette technique pour l'extraction d'événements.

Résultats avec repérage automatique des entités nommées Pour compléter les résultats précédents basés sur une annotation manuelle des entités nommées, nous avons évalué notre approche mixte avec une annotation automatique des entités. Nous observons une baisse globale des performances des trois approches et plus particulièrement du taux de rappel bien que les performances restent acceptables (64% de F-mesure pour l'approche mixte). Ce point est important car dans une application réelle d'extraction d'événements, les entités nommées sont toujours repérées par des outils d'extraction automatique.

Améliorations Pour améliorer la combinaison, nous envisageons de tester plusieurs techniques d'hybridation. Tout d'abord, afin d'obtenir de meilleurs résultats de façon plus globale (c'est-à-dire maximiser la F-mesure), nous expérimenterons l'ajout d'un système d'estimation de confiance au sein de nos extracteurs. Cela peut être réalisé par différents moyens : (1) faire évaluer à la main par un expert linguiste chaque règle/motif composant les deux approches précédentes et reporter cette confiance sur les événements/relations extraits ; (2) estimer la confiance de chaque règle/motif automatiquement en réalisant une évaluation préalable. Une dernière piste à explorer est l'apport des approches statistiques : nous souhaitons apprendre automatiquement un modèle de performances qui permettrait une sélection contextuelle d'approche en suggérant, lors du traitement d'un corpus, la meilleure approche à utiliser.

6 Conclusion et perspectives

Dans cet article nous avons proposé une étude comparative de deux approches et de leur combinaison pour l'extraction automatique d'événements. Nos résultats montrent que la méthode

mixte améliore significativement la qualité des événements extraits. Malgré une combinaison plutôt simple, ces résultats sont encourageants et nous invitent à explorer de nouveaux modes d'hybridation afin de tirer le meilleur parti des deux premières approches (améliorer le taux de rappel sans perdre trop en précision).

Références

AGRAWAL, R., IMIELIŃSKI, T. et SWAMI, A. (1993). Mining association rules between sets of items in large databases. SIGMOD '93, New York. ACM.

AHN, D. (2006). The stages of event extraction. ARTE '06, pages 1–8, Stroudsburg, USA. ACL.

AONE, C. et RAMOS-SANTACRUZ, M. (2000). Rees : A large-scale relation and event extraction system. *In ANLP*, pages 76–83.

BÉCHET, N., CELLIER, P., CHARNOIS, T. et CRÉMILLEUX, B. (2012). Discovering linguistic patterns using sequence mining. *In CICLing (1)*, pages 154–165.

CAPET, P., DELAVALLADE, T., GÉNÉREUX, M., POIBEAU, T., SÁNDOR, Á. et VOYATZI, S. (2011). Un système de détection de crise basé sur l'extraction automatique d'événements. *In* et P. HOOGSTOEL, M. C., éditeur : *Sémantique et multimodalité en analyse de l'information*, pages 293–313. Lavoisier.

CELLIER, P. et CHARNOIS, T. (2010). Fouille de données séquentielle d'itemsets pour l'apprentissage de patrons linguistiques. *In TALN (short paper)*.

DAVIDSON, D. (2001). *Essays on Actions and Events*. Oxford University Press.

DONG, G. et PEI, J. (2007). *Sequence Data Mining*. Advances in Database Systems. Kluwer.

FERRÉ, S. (2009). Camelis : a logical information system to organize and browse a collection of documents. *In Int. J. General Systems*, volume 38.

GRISHMAN, R., HUTTUNEN, S. et YANGARBER, R. (2002). Information extraction for enhanced access to disease outbreak reports. *Journal of Biomedical Informatics*, 35(4):236–246.

KRIEG-PLANQUE, A. (2009). *A propos des noms propres d'événement*, volume 11, pages 77–90. Les carnets du Cediscor.

NADEAU, D. et SEKINE, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26. Publisher : John Benjamins Publishing Company.

POIBEAU, T. (2003). *Extraction automatique d'information : Du texte brut au web sémantique*. Lavoisier.

SAVAL, A., BOUZID, M. et BRUNESSAUX, S. (2009). A semantic extension for event modelisation. *21st IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2009)*.

SERRANO, L., BOUZID, M., CHARNOIS, T. et GRILHERES, B. (2012). Vers un système de capitalisation des connaissances : extraction d'événements par combinaison de plusieurs approches. *In SOS-DLWD'2012 at EGC'2012*.

SERRANO, L., GRILHERES, B., BOUZID, M. et CHARNOIS, T. (2011). Extraction de connaissances pour le renseignement en sources ouvertes. *In SOS'2011 at EGC'2011*.

VAN DE VELDE, D. (2006). *Grammaire des événements*. Presses Universitaires du Septentrion.

XU, F., USZKOREIT, H. et LI, H. (2006). Automatic event and relation detection with seeds of varying complexity. *In AAAI Workshop Event Extraction and Synthesis*, Boston.