Extraction et agrégation automatique d'événements pour la veille en sources ouvertes : du texte à la connaissance

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

Extraction et agrégation automatique d'événements pour la veille en sources ouvertes : du texte à la connaissance

Laurie Serrano[1,2], Maroua Bouzid[1], Thierry Charnois[1], Stephan Brunessaux[2], Bruno Grilheres[2]

[1] GREYC, Université de Caen Basse-Normandie
Campus Côte de Nacre, Boulevard du Maréchal Juin
BP 5186 - 14032 Caen
prenom.nom@unicaen.fr
[2] Département IPCC, Cassidian
Parc d'Affaires des Portes - 27600 Val de Reuil
prenom.nom@cassidian.com

Résumé :
Le travail présenté ici est réalisé dans le cadre de recherches en cours visant à développer un système global de capitalisation des connaissances. Le but final de ce système est de construire des fiches de connaissances résumant tout le savoir acquis à propos d'événements extraits à partir de textes. L'extraction de ces événements est réalisée de façon automatique grâce à notre outil fondé sur deux approches actuelles d'extraction d'information. Dans une seconde phase, nous proposons des mécanismes d'agrégation en tant que post-traitement nécessaire pour convertir les résultats d'extraction en réelle connaissance.

Mots-clés : Capitalisation des connaissances, extraction d'événements, ontologies, Web sémantique, veille en sources ouvertes

1 Introduction

Même s'il existe actuellement de nombreux systèmes de gestion de contenu et outils de veille professionnelle, les analystes du renseignement d'origine source ouverte (ROSO) [1] recueillent encore manuellement une

1. Collecte de l'intelligence à partir de sources disponibles librement et publiquement comme les journaux, la radio, la télévision ou Internet

grande partie des informations qui les intéressent. Généralement, cette collecte vise à créer des fiches de connaissances résumant tout le savoir (faits, propriétés, etc.) acquis à propos d'une entité (personne, produit, événement, etc.). En réponse à cela, le département IPCC (Information Processing, Control and Cognition) développe le WebLab [2], une plateforme open-source de "media mining" dédiée à l'intégration de divers outils pour faciliter la découverte de connaissances et la prise de décision. Dans ce cadre, nous nous intéressons à l'extraction automatique d'événements dans des dépêches de presse par combinaison de plusieurs approches et à des mécanismes d'agrégation pour améliorer la qualité des fiches créées et des connaissances capitalisées.

La fouille de textes est un domaine de recherche récent visant à analyser automatiquement le langage naturel dans le but d'en extraire des informations d'intérêt (Hobbs & Riloff, 2010). Les tâches les plus communes sont la reconnaissance d'entités nommées (Nadeau & Sekine, 2007), l'extraction de relations entre ces entités (Rosario & Hearst, 2005) et la détection d'événements. Cette dernière tâche est particulièrement utile aux veilles stratégique et économique. Deux types d'approches sont traditionnellement distinguées : les méthodes symboliques fondées sur des règles linguistiques construites manuellement (Grishman & Sundheim, 1996) et les méthodes d'apprentissage supervisé à base de modèle numérique (Ahn, 2006). Les systèmes purement linguistiques, bien que généralement très précis, ont pour principales faiblesses un taux de rappel faible et un coût de développement manuel élevé. Les approches statistiques permettent de couvrir de nombreux contextes d'apparition mais nécessitent une grande quantité de données annotées et produisent un modèle de type "boite noire" difficilement accessible et modifiable. Par ailleurs, les méthodes d'apprentissage de règles linguistiques ou les approches semi-supervisées apparaissent intéressantes car elles visent à tirer le meilleur parti des deux approches ci-dessus (Xu *et al.*, 2006).

Ces techniques utilisées seules restant imparfaites, nous proposons de les combiner et d'améliorer la qualité globale des résultats grâce à des méthodes de fusion de données. La fusion de données textuelles a été l'objet de nombreux travaux dans plusieurs domaines et particulièrement dans la communauté des bases de données à travers les recherches en "record linkage" ou "duplicate record detection". (Winkler, 2006) propose un état de l'art de ces travaux qui procèdent généralement en deux étapes : une phase dite de préparation des données suivie d'une phase de fusion des champs

2. `http://weblab-project.org`

référant à une même entité. Nous proposons d'adapter ces travaux à l'agrégation d'événements automatiquement extraits à partir de textes. En effet, les outils d'extraction actuels ne produisent pas de la connaissance mais retournent des informations parcellaires qui peuvent être imparfaites, incomplètes ou redondantes. Par conséquent, dans le but de construire des fiches de connaissances, nous devons élaborer un système plus complet de découverte de connaissances combinant des techniques d'extraction et d'agrégation d'informations. Cet article adresse ce problème et est structuré comme suit : la section 2 présente notre modélisation des connaissances centrée sur la représentation des événements et de leurs dimensions. Puis, la section 3 présente le fonctionnement et l'évaluation de notre extracteur d'événements. La section 4 décrit notre méthode d'agrégation combinant différentes mesures de similarité spécifiques à chaque dimension. Enfin, la section 5 conclue cet article et aborde nos futurs travaux.

2 Un modèle d'événement

Les événements sont les entités centrales de notre système de capitalisation des connaissances. Le concept d'"événement" a initialement été étudié en philosophie (Davidson, 1980) puis en linguistique (Desclès, 1990). On distingue actuellement deux modèles largement utilisés. D'une part, l'approche TimeML (Pustejovsky *et al.*, 2003) définit un événement comme tout terme temporellement ancré et est généralement choisi dans des projets visant la construction de chronologies. D'autre part, dans le modèle ACE (NIST, 2005) un événement est vu comme une structure complexe impliquant plusieurs arguments. Ce modèle est focalisé sur un ensemble pré-défini de types et sous-types d'événement (e.g. Life, Movement, Business, etc.) et associe à chaque événement un ensemble d'arguments (Time, Place, Agent, Instrument, etc.). Même s'il nécessite des adaptations, ce modèle, plus riche et tenant compte de l'intérêt des événements pour une application donnée, convient mieux à nos besoins de modélisation.

Afin de proposer une définition formelle des événements, nous nous fondons également sur les travaux de (Saval *et al.*, 2009) décrivant une extension sémantique pour la modélisation d'événements de type catastrophes naturelles. Les auteurs définissent un événement E comme la combinaison d'une propriété sémantique S, d'un intervalle temporel T et d'une entité spatiale SP. Nous adaptons cette modélisation à notre problématique en y ajoutant une quatrième dimension A pour représenter les participants impliqués dans un événement et leurs rôles respectifs. Par consé-

quent, un événement est représenté comme suit :

Définition 1
Un événement E est modélisé comme $E < S, T, SP, A >$ où la propriété sémantique S est le type de l'événement, l'intervalle temporel T est la date à laquelle l'événement est survenu, l'entité spatiale SP est le lieu d'occurrence de l'événement et A est l'ensemble des participants impliqués dans E associés avec le(s) rôle(s) qu'ils tiennent dans E.

Exemple 1
L'événement exprimé par "M. Dupont a mangé au restaurant Lafayette à Paris en 1999" est représenté comme (Manger, 1999, Paris, M. Dupont).

Nous avons également observé la façon dont les événements sont exprimés dans les dépêches de presse (e.g. celles de l'AFP) : de façon générale, un événement principal est reporté, résumé dans le titre et développé tout au long de la dépêche (en faisant parfois référence à d'autres événements). Cette description est constituée de plusieurs sous-événements ("mentions d'événement" dans le modèle ACE) qui sont généralement composés d'un nom d'événement ("ancre") relié à une ou plusieurs autres entités telles qu'une date, un lieu et des participants ("arguments"). Notre objectif principal est d'extraire automatiquement les mentions d'événement pertinentes pour notre application et ensuite d'agréger celles qui réfèrent à un seul et même événement dans la réalité. Les sections suivantes décrivent comment chaque dimension de l'événement est modélisée.

2.1 Dimension sémantique

Pour définir précisément quelles sont les informations d'intérêt pour notre application, nous avons développé une ontologie de domaine constituant la base de notre système de capitalisation des connaissances. Cette ontologie est implémentée au format OWL [3] (selon les recommandations du W3C [4] pour la représentation des connaissances au sein du Web Sémantique) et a été élaborée suite à un état de l'art des ontologies et modélisations existantes. Nous avons étudié des ontologies dites de "haut niveau" mais aussi des spécifications dédiées au ROSO telles que les standards OTAN (NATO, 2005). Inspirés de ces travaux, nous avons construit un

3. Ontology Web Language, `http://www.w3.org/TR/owl-features/`
4. World Wide Web Consortium, `http://www.w3.org/`

modèle de connaissances en prenant soin de maintenir des équivalences sémantiques avec les représentations existantes.

L'interprétation d'un événement dépend étroitement de la sémantique exprimée par les termes employés pour nommer cet événement. Ces noms d'événement transposent en langage naturel la propriété sémantique des événements évoquée par (Saval *et al.*, 2009). La dimension S d'un événement représente le type de l'événement, c'est-à-dire sa classe conceptuelle au sein de notre ontologie de domaine. Dans notre application, existent environ 20 types d'événement-cible, organisés en taxonomie et principalement liés au domaine militaire et de la sécurité (attaques, enlèvements, événements nucléaires, trafics illégaux, etc.).

2.2 Dimension temporelle

La manipulation d'entités temporelles extraites de textes nécessite généralement une phase préalable de normalisation de par les multiples façons d'exprimer le temps en langage naturel ("04/09/2012", "Mon, 09/April/12", "two days ago", etc.). Pour cela, nous utilisons le "Time Unit System" (TUS) proposé par (Ladkin, 1987). Il s'agit d'une approche hiérarchique et granulaire qui représente toute expression temporelle en un groupe de granules (c'est-à-dire des unités temporelles indivisibles). Un granule (ou unité de temps) est une séquence finie d'entiers organisés selon une hiérarchie linéaire : année, mois, jour, heure, etc. De plus, ce formalisme introduit la notion de BTU (Basic Time Unit) qui correspond au niveau de granularité choisi en fonction de la précision nécessitée par une application (e.g. les jours, les secondes, etc.). Par exemple, si le BTU est fixé à *heure*, chaque unité temporelle sera exprimée comme une séquence d'entiers i telle que : *i = [année,mois,jour,heure]*. De plus, TUS définit la fonction $max_j([a_1, a_2, ..., a_{j-1}])$ donnant la valeur maximale possible à la position j pour qu'une séquence temporelle soit valide en tant que date. Cet opérateur est nécessaire car, selon notre actuel système calendaire, le granule *jour* dépend des granules *mois* et *année*.

Pour notre application, nous choisissons un BTU *jour* correspondant à la précision maximale des dates extraites. Par conséquent, toute expression temporelle i aura la forme suivante : *i = [année,mois,jour]*. Par exemple, [2010,09,19] représente un intervalle de temps qui débute le 18 septembre 2012 à minuit et termine un jour plus tard (ce qui équivaut à un BTU).

Enfin, les entités temporelles extraites peuvent s'avérer plus ou moins précises. Dans certains cas, les expressions de temps peuvent être imprécises à l'origine (e.g. "en Mai 2010") et, dans d'autres cas, l'imprécision

peut être causée par une erreur d'extraction. Pour représenter ces entités floues, nous introduisons le symbole $\emptyset$ défini comme le manque d'information au sens général. Soit $T = [g_1, g_2, g_3]$ une expression temporelle :

Définition 2
T est une date dite : complète si $\forall i \in \{1, 2, 3\}, g_i \neq \emptyset$*, incomplète si* $\exists i \in \{1, 2, 3\}, g_i = \emptyset$

2.3 Dimension spatiale

Nous représentons les entités spatiales comme des aires géographiques et utilisons les relations topologiques du modèle RCC-8 pour leur agrégation (Cohn & Hazarika, 2001) : "disconnected" (DC), "externally connected" (EC), "equal" (EQ), "partially overlapping" (PO), "tangential proper part" (TPP), "tangential proper part inverse" (TPPi), "non-tangential proper part" (NTPP), "non-tangential proper part inverse" (NTPPi).

Comme dans le cas des entités temporelles, le raisonnement spatial nécessite d'opérer sur des objets non-ambigus et nous devons par conséquent préciser géographiquement tous les lieux extraits par notre outil. Dans le cadre du WebLab, nous nous intéressons notamment à la désambiguïsation d'entités spatiales dans le but d'associer un identifiant GeoNames [5] unique (une URI) à chaque lieu extrait et d'effectuer des traitements plus avancés comme la géolocalisation ou l'inférence spatiale. Utiliser une base géographique comme GeoNames a plusieurs avantages : tout d'abord, il s'agit d'une base open-source et sémantique, par conséquent bien adaptée à une intégration au sein du WebLab ; de plus, en complément des coordonnées géographiques, cette ressource fournit des relations topologiques entre lieux, comme par exemple des relations d'inclusion. Nous réutilisons les trois propriétés suivantes pour agréger les événements (cf.section 4.3) :

- la propriété "children" réfère à une inclusion administrative ou physique entre deux entités géographiques ;
- la propriété "nearby" relie deux entités qui sont géographiquement proches l'une de l'autre ;
- la propriété "neighbour" est utilisée lorsque deux entités géographiques partagent au moins une frontière.

5. `http://www.geonames.org/`

2.4 Dimension agentive

Comme nous l'avons dit précédemment, tous les participants d'un événement et leurs rôles respectifs sont représentés formellement par la composante A. Nous définissons cette dimension de l'événement comme un ensemble $A = (P_i, r_j)$ où chaque élément est un couple composé d'un participant p_i et d'un rôle r_j et où i et $j \in \mathrm{N}$. Notre modèle ne limite pas la nature du champ "participant" (chaîne de caractères, entité nommée, nom propre/commun, etc.) pour rester le plus générique possible. Toutefois, dans notre application un participant correspond concrètement à une entité nommée de type "Personne" ou "Organisation" ayant été extraite et liée automatiquement à l'événement dans lequel elle est impliquée. Les rôles des participants n'étant actuellement pas extraits par notre système, cet aspect ne sera pas traité ici.

3 Extraction automatique d'événements

Notre première contribution est le développement d'un outil d'extraction automatique visant à repérer l'ensemble des événements défini en section 2.1 et dédié au traitement de dépêches de presse en langue anglaise. Nous avons développé deux extracteurs suivant deux approches distinctes : une méthode symbolique constituée de règles linguistiques élaborées manuellement et une approche par apprentissage de patrons linguistiques.

Le premier extracteur a été implémenté grâce à la plateforme GATE [6] et repose sur une chaîne de traitement composée de différents modules d'analyse linguistique ("tokenisation", découpage en phrases, repérage lexical, étiquetage grammatical, analyse syntaxique, etc.). Nous définissons tout d'abord un ensemble de termes considérés comme possibles déclencheurs d'événement ("ancres 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 manuellement répartis en différentes listes, chacune étant associée à un type d'événement (c'est-à-dire à une classe de notre ontologie) afin d'être repérés et annotés dans le corpus à analyser. Nous associons ensuite à ces ancres d'événement les différentes entités qu'elles impliquent. Pour cela, nous effectuons, dans un premier temps, une extraction automatique d'en-

6. General Architecture for Text Engineering, `http://gate.ac.uk/`

tités nommées [7] grâce à un autre module GATE que nous avons développé (Serrano *et al.*, 2011) ainsi qu'une analyse en constituants syntaxiques (syntagmes verbaux, nominaux, etc.). Enfin, ces entités sont liées à l'ancre d'événement appropriée grâce à une analyse syntaxique en dépendance (réalisée par le Stanford Parser [8]) ainsi que des règles de grammaire élaborées manuellement. Finalement, nous obtenons une annotation positionnée sur l'ancre d'événement indiquant le type de l'événement ainsi que les différentes entités impliquées (date, lieu et participants).

Dans un second temps, nous nous sommes intéressés 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 (Béchet *et al.*, 2012) à l'extraction d'information dans les textes. Ceux-ci s'intéressent en particulier à l'extraction de motifs séquentiels d'"itemsets", c'est-à-dire repérer, dans un ensemble de séquences, des enchaînements 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'items et d'obtenir ainsi des patrons génériques, spécifiques ou mixant les informations (ce qui n'est pas permis par les motifs d'items simples). Par exemple, cette technique permet d'extraire les patrons suivants : <*trois chiliens arrêtés près de Buenos Aires*> <*SN arrêtés près de Lieu*> <*SN VB PRP Location*> [9], 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 d'abord, l'annotation manuelle de corpus reste un effort important et l'analyse syntaxique est

7. Nous repérons les dates, lieux, noms de personnes et d'organisations

8. `http://nlp.stanford.edu/software/lex-parser.shtml`

9. SN : syntagme nominal, VB : verbe, PRP : préposition, Lieu : entité nommée de type "lieu"

encore une technologie aux performances inégales et peu disponible librement selon les langues. Le point faible des méthodes de fouille reste le nombre important de motifs extraits. Pour pallier ce problème, (Béchet *et al.*, 2012) propose l'ajout de contraintes pour diminuer la quantité de motifs retournés et l'utilisation de l'outil Camelis (Ferré, 2009) pour ordonner et visualiser les motifs des plus généraux aux plus spécifiques puis filtrer les plus pertinents.

Lors d'une première évaluation (Serrano *et al.*, 2012), nous avons évalué chaque extracteur séparément, puis mesuré la qualité des événements obtenus en opérant une simple union de leurs résultats. Cette union augmente significativement la qualité des événements extraits : nous obtenons un F1-score de 70 % (90 % de précision et 60 % de rappel), ce qui équivaut à 10 points de plus que la meilleure des deux approches évaluée séparément. Un examen plus fin des résultats a montré que l'apprentissage de patrons linguistiques complète avec succès le premier extracteur symbolique en améliorant le rappel, toutefois, la simplicité d'une union entraîne une légère perte de précision. Cette première évaluation confirme le besoin d'un processus de fusion plus élaboré qui permettra de tirer le meilleur parti des deux approches. Nous présentons ci-après un processus d'agrégation basé sur plusieurs mesures de similarité entre événements extraits.

4 Similarité entre événements

L'extraction d'information étant encore à un stade de recherche, les outils actuels fournissent des résultats pouvant être incomplets, redondants, flous, conflictuels, imprécis et parfois totalement erronés. Comme nous l'avons montré plusieurs approches s'avèrent complémentaires si leurs résultats sont combinés de façon adaptée.

Notre objectif principal est d'améliorer la qualité de l'extraction des événements en agrégeant les mentions d'événement provenant des deux extracteurs. Pour cela, nous proposons d'évaluer la similarité entre événements c'est-à-dire estimer avec quel degré deux mentions d'événement peuvent référer à un seul et même événement de la réalité. Dans notre application, cette similarité permettra à l'utilisateur final de compléter ses connaissances et de décider, le cas échéant, de fusionner deux événements et donc deux fiches de connaissances (une interface d'aide à la fusion de fiches a été développée au sein de la plateforme WebLab). Ci-après, nous détaillons les mesures de similarité utilisées pour chacune des dimensions de l'événement et les exprimons selon une échelle qualitative. Nous défi-

nissons cette échelle comme composée de quatre niveaux :

1. identité (ID) : les deux dimensions réfèrent à la même entité réelle
2. similarité (Sim) : il y a de l'imprécision dans l'une ou les deux dimensions mais elles ont des caractéristiques en commun et pourraient référer à la même entité du monde réel
3. compatibilité (Co) : il y a un manque d'information dans l'une ou les deux dimensions qui empêche de dire si elles réfèrent ou non à la même entité réelle
4. incompatibilité (INC) : les deux dimensions ne réfèrent pas à la même entité dans la réalité

Définition 3
Soient $e = (S, T, SP, A)$ *et* $e' = (S', T', SP', A')$ *deux événements tels que* $e, e' \in \mathcal{E}$ *(*$\mathcal{E}$ *étant l'ensemble des événements extraits), tout type de fonction de similarité* $R(e, e')$ *a pour domaine de définition :*

$$R : \mathcal{E} \times \mathcal{E} \rightarrow \{ID, Sim, Co, INC\}$$

Ces niveaux sont définis et illustrés dans les sections suivantes, à travers différentes fonctions de similarité spécifiques à chaque dimension d'événement : R_s, R_t, R_{sp} et R_a.

Soulignons ici que le niveau "Sim" peut ne pas être défini selon la nature de certaines dimensions. En effet, la représentation temporelle que nous considérons est universelle, standard et indépendante de toute interprétation. Pour cette raison, notre définition de la similarité entre deux entités temporelles ne considère que trois niveaux : l'identité, la compatibilité et l'incompatibilité (cf. section 4.2). Cependant, la dimension spatiale peut être interprétée différemment selon le niveau d'abstraction, le point de vue de l'observateur ou encore la référence utilisée. Cette spécificité de l'information spatiale nous amène à définir non seulement l'identité, la compatibilité et l'incompatibilité mais aussi la similarité exprimant les cas où deux entités spatiales exprimées selon deux points de vue différents peuvent référer à une même entité réelle (cf. section 4.3).

4.1 Similarité sémantique

Définition 4
Nous définissons R_s *une fonction de similarité sémantique telle que :*

$$R_s : \mathcal{E} \times \mathcal{E} \rightarrow \{ID, Co, INC\}$$

Soient $e = (S, T, SP, A)$ et $e' = (S', T', SP', A')$ deux événements, nous définissons la similarité sémantique de la façon suivante :

Définition 5
- $R_s(e, e') = INC$ *ssi* S *et* S' *sont deux classes distinctes de l'ontologie et ne sont pas en relation de subsomption*
- $R_s(e, e') = Co$ *ssi* S *est une sous-classe (à tout niveau) de* S' *dans l'ontologie (et inversement)*
- $R_s(e, e') = ID$ *ssi* S *et* S' *sont une même classe dans l'ontologie*

Le niveau "Sim" n'est pas défini ici car ces trois niveaux suffisent pour notre application mais nous envisageons de poursuivre ce travail en étudiant des mesures de similarité plus avancées (notamment en explorant les recherches en alignement d'ontologies) qui permettraient l'agrégation de plusieurs événements en des entités plus larges tels que des phénomènes.

4.2 Similarité temporelle

Définition 6
Nous définissons R_t *une fonction de similarité temporelle telle que :*

$$R_t : \mathcal{E} \times \mathcal{E} \rightarrow \{ID, Co, INC\}$$

Soient $T = [g_1, g_2, g_3]$ et $T' = [g'_1, g'_2, g'_3]$ deux expressions temporelles converties au format TUS.

Définition 7
Nous définissons $\oplus$ *un opérateur de complétion prenant comme paramètres* T *et* T' *où* T *ou* T' *est incomplet (cf. section 2.2) et retournant une expression temporelle* $T'' = [g''_1, g''_2, g''_3]$ *telle que :*
- *si* $g_i = g'_i$ *alors* $g''_i = g_i = g'_i$
- *si* $g_i = \emptyset$ *alors* $g''_i = g'_i$ *(et inversement)*

Soient $e = (S, T, SP, A)$ et $e' = (S', T', SP', A')$ deux événements, nous définissons la similarité temporelle comme suit :

Définition 8
- $R_t(e, e') = INC$ *ssi*
  - $\exists_{i \in \{1,2,3\}}$ *tel que* $g_i \neq \emptyset$ *et* $g'_i \neq \emptyset$ *et* $g_i \neq g'_i$
  - *si* T *ou* T' *est incomplet alors* $T \oplus T'$ *retourne une date invalide (cf. section 2.2)*
- $R_t(e, e') = Co$ *ssi* T *ou* T' *est incomplet et* $T \oplus T'$ *retourne une date valide (cf. section 2.2)*
- $R_t(e, e') = ID$ *ssi* T *et* T' *sont complets et* $\forall_{i \in \{1,2,3\}}$, $g_i = g'_i$

4.3 Similarité spatiale

Définition 9
Nous définissons R_{sp} une fonction de similarité spatiale telle que :

$$R_{sp} : \mathcal{E} \times \mathcal{E} \rightarrow \{ID, Sim, Co, INC\}$$

Pour estimer la similarité entre entités spatiales nous utilisons le modèle RCC-8 (cf. section 2.3) ainsi que les relations topologiques de la base GeoNames. La dimension spatiale d'un événement peut varier avec le point de vue à partir duquel celui-ci est reporté : par exemple, un événement survenu à Cestas (une petite ville à côté de Bordeaux) sera précisément situé par une dépêche française mais pourrait être plus globalement localisé à Bordeaux (la grande ville la plus proche) par un journal étranger. Cette similarité peut également être influencée par la nature des deux entités comparées (ville, pays, quartier, etc.) et par la taille de l'entité administrative les englobant (e.g. la distance entre deux villes au Liechtenstein ne sera pas évaluée de la même façon que pour deux villes en Russie). Pour résumer, l'absence de relation topologique entre deux lieux peut être due à une différence de point de vue mais ne signifie pas forcément une incompatibilité spatiale. Pour traiter au mieux ces cas, nous introduisons $D_{sp}(SP, SP')$ une fonction générique retournant une distance spatiale entre deux entités géographiques et un seuil d_{sp} pour discriminer incompatibilité et compatibilité spatiale. D_{str} et d_{sp} seront fixés lors de nos futures expérimentations.

Soient $e = (S, T, SP, A)$ et $e' = (S', T', SP', A')$ deux événements, nous définissons la similarité spatiale comme suit :

Définition 10
Soit $r(SP, SP')$ la relation RCC-8 liant les deux entités spatiales :

- $R_{sp}(e, e') = INC$ *ssi* $r(SP, SP') = DC$ *et* $D_{sp}(SP, SP') > d_{sp}$ *(SP et SP' ne sont liés par aucune relation dans GeoNames et $D_{sp}(SP, SP') > d_{sp}$)*
- $R_{sp}(e, e') = Co$ *ssi* $r(SP, SP') = DC$ *et* $D_{sp}(SP, SP') < d_{sp}$ *ou SP ou SP' est inconnu (SP et SP' ne sont liés par aucune relation dans GeoNames et $D_{sp}(SP, SP') < d_{sp}$ ou $SP \equiv \emptyset$ ou $SP' \equiv \emptyset$)*
- $R_{sp}(e, e') = Sim$ *ssi* $r(SP, SP') \in \{EC, PO, TPP, NTPP, TPPi, NTPPi\}$ *(SP et SP' sont liés dans GeoNames par une relation de type "nearby", "neighbour" ou "children")*
- $R_{sp}(e, e') = ID$ *ssi* $r(SP, SP') = EQ$ *(SP' ont la même URI dans GeoNames)*

4.4 Similarité agentive

Définition 11
Nous définissons R_a une fonction de similarité agentive telle que :

$$R_a : \mathcal{E} \times \mathcal{E} \rightarrow \{ID, Sim, Co, INC\}$$

Comme mentionné précédemment, l'ensemble des participants d'un événement consiste en une ou plusieurs entités nommées de type "Personne" ou "Organisation" stockées dans la base de connaissances sous forme de chaînes de caractère. Par conséquent, l'agrégation pour cette dimension implique l'utilisation de mesures de similarité dédiées aux chaînes de caractère mais également adaptées à la comparaison de telles entités. Pour cela, nous avons exploré les méthodes actuelles et plus particulièrement celles retenues par (Elmagarmid *et al.*, 2007). Pour notre application, la distance de Jaro (Jaro, 1976) semble être un bon choix de par son temps d'exécution modéré et sa bonne capacité à traiter la similarité entre noms de personne. Notre approche visant à rester indépendante des mesures de similarité choisies pour chaque dimension, nous prévoyons d'évaluer différentes métriques au sein de la plateforme WebLab pour motiver notre choix final et définir un seuil de distance d_{str} utilisé par la suite.

Nous définissons $D_{str}(P, P')$ comme une fonction générique retournant une distance entre deux chaînes de caractère données telle que $D_{str}(P, P') = 0$ signifie que P et P' sont identiques. De plus, comme nous traitons d'entités nommées et non seulement de chaînes de caractère, nous utilisons la base sémantique DBpedia [10] pour obtenir les différents noms alternatifs existants pour référer à une même entité du monde réel. Nous définissons $alt(P)$ comme une fonction donnant tous les noms alternatifs présents dans DBpedia pour un participant donné P. Cette technique pourra être améliorée en explorant d'autres techniques de résolution de coréférence entre entités nommées (Finin *et al.*, 2009).

Définition 12
$P = P'$ *ssi* $P \in alt(P')$ *(et inversement) ou* $D_{str}(P, P') = 0$

De plus, comme défini en section 2.4, la dimension agentive correspond à un ensemble de participants, l'agrégation pour cette dimension doit donc pouvoir gérer une similarité entre ensembles d'entités.

Définition 13
$A = A'$ *ssi* $|A| = |A'|$ *et* $\forall_{P \in A}$, $\exists_{P' \in A'}$ *tel que* $P = P'$ *(et inversement)*

10. `http://dbpedia.org/`

Nous définissons quatre niveaux de similarité agentive comme suit :

Définition 14
Soient $e = (S, T, SP, A)$ *et* $e' = (S', T', SP', A')$ *deux événements tels que* $A = \{P_1, P_2, \ldots, P_n\}$ *et* $A' = \{P'_1, P'_2, \ldots, P'_n\}$ *:*

- $R_a(e, e') = INC$ *ssi* $\forall_{P \in A}$ *et* $\forall_{P' \in A'}$, $P \notin alt(P')$ *(et inversement) et* $D_{str}(P, P') < d_{str}$
- $R_a(e, e') = Co$ *ssi* $\forall_{P \in A}$, $\exists_{P' \in A'}$ *tel que* $D_{str}(P, P') \geq d_{str}$ *ou ssi* $|A| = 0$ *ou* $|A'| = 0$
- $R_a(e, e') = Sim$ *ssi* $\forall_{P \in A}$, $\exists_{P' \in A'}$ *tel que* $P = P'$
- $R_a(e, e') = ID$ *ssi* $A = A'$

4.5 Agrégation automatique d'événements

Il est communément admis dans les applications de veille en sources ouvertes que l'information manipulée par les analystes est incertaine à plusieurs niveaux : l'information en elle-même, la source, les traitements opérés, etc. Partant de ce constat, nous voulons laisser à l'utilisateur le choix final de fusionner (ou non) deux événements jugés similaires, et cela dans le but d'éviter tout perte d'information pouvant survenir avec une fusion totalement automatisée. L'objectif de nos travaux est donc d'agréger les événements selon leur degré de similarité. Pour cela, nous organisons l'ensemble des événements extraits E en un graphe $G =< E, edge >$ où chaque nœud est un événement et chaque arc un degré de similarité. La construction de ce graphe est faite comme suit : chaque événement candidat e_c provenant du système d'extraction est comparé à chaque nœud e du graphe existant G (c'est-à-dire à chaque événement déjà stocké). Cette comparaison "deux à deux" consiste à estimer les niveaux de similarité de chaque dimension (R_s, R_t, R_{sp} et R_a). Si e_c est globalement identique (c'est-à-dire que les quatre types de similarité équivalent à "ID") à au moins un événement e du graphe alors e_c n'est pas inséré dans celui-ci. Dans le cas contraire, nous créons un nouveau nœud dans le graphe G correspondant à e_c et les arcs reliant e_c à chaque événement e de l'ensemble E indiquant les niveaux de similarité estimés pour chaque dimension (nous obtenons donc quatre types d'arc par paire d'événements). Au sein de la plateforme WebLab, la connaissance provenant des différents services de traitement de l'information est stockée dans une base sémantique. Toute cette connaissance ainsi que le graphe résultant de notre processus d'agrégation est rendu accessible à l'utilisateur au travers de diverses interfaces (recherche, vue d'un document annoté, carte géographique, bandeau tem-

porel et autres vues). Les niveaux de similarité que nous avons estimés sont proposés comme critères supplémentaires lors de la recherche d'événements au sein du WebLab : par exemple, l'utilisateur peut demander au système de lui renvoyer tous les événements jugés similaires (niveau "Sim") et ensuite examiner les fiches de connaissances correspondantes pour décider (selon ses propres connaissances) quels événements sont véritablement identiques dans le monde réel.

5 Conclusions et perspectives

Nous avons proposé une modélisation de l'événement en quatre dimensions (sémantique, temporelle, spatiale et agentive), puis, un système d'extraction d'événements fondé sur deux approches : une méthode linguistique reposant sur une analyse syntaxique en dépendance et des règles d'extraction élaborées manuellement et une technique de découverte de motifs visant à extraire automatiquement les patrons linguistiques les plus fréquents d'un ensemble de textes. Notre évaluation a montré que ces approches ont des forces complémentaires et qu'un processus d'agrégation adapté permettra d'améliorer la qualité de la connaissance acquise. Nous proposons donc un processus d'agrégation d'événements basé sur différentes mesures de similarité entre événements spécifiques à chaque dimension et exprimées selon une échelle qualitative. Une évaluation complète de notre système capitalisation des connaissances est en cours de réalisation. Le travail présenté est centré sur la définition d'un système global de reconnaissance et d'agrégation d'événements le plus générique possible. Cela implique que les différentes mesures de similarité présentées soient plutôt simples et facilement interchangeables avec d'autres mesures plus avancées sans remettre en cause notre approche globale. La similarité agentive pourra, par exemple, être améliorée en y intégrant des techniques de résolution de coréférence plus abouties. Nous pourrons également prendre en compte certaines dépendances entre dimensions d'événement : à titre d'exemple, la dimension sémantique peut influencer l'agrégation d'entités temporelles dans le cas d'événements duratifs comme des épidémies ou des guerres où deux mentions d'événement peuvent avoir deux dates différentes mais tout de même référer au même événement du monde réel.

Références

AGRAWAL R., IMIELIŃSKI T. & SWAMI A. (1993). Mining association rules between sets of items in large databases. In *SIGMOD'93*, p. 207–216.

AHN D. (2006). The stages of event extraction. In *ARTE '06*, p. 1–8.

BÉCHET N. *et al.* (2012). Discovering linguistic patterns using sequence mining. In *CICLing (1)*, p. 154–165.

COHN A. G. & HAZARIKA S. M. (2001). Qualitative spatial representation and reasoning : An overview. *Fundam. Inf.*, **46**, 1–29.

DAVIDSON D. (1980). *Essays on Actions and Events*. Oxford University Press.

DESCLÈS J.-P. (1990). "State, event, process and topology". *General Linguistics*, **29**(3), 159–200.

ELMAGARMID A. K., IPEIROTIS P. G. & VERYKIOS V. S. (2007). Duplicate record detection : A survey. *IEEE TKDE*, **19**(1), 1–16.

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

FININ T. *et al.* (2009). Using wikitology for cross-document entity coreference resolution. In *AAAI'09*, p. 29–35.

GRISHMAN R. & SUNDHEIM B. (1996). Message understanding conference-6 : a brief history. In *ACL'96*, p. 466–471.

HOBBS J. R. & RILOFF E. (2010). Information extraction. In *Handbook of Natural Language Processing, 2nd Edition*. CRC Press.

JARO M. A. (1976). *UNIMATCH : A Record Linkage System : User's Manual*. Rapport interne, U.S. Bureau of the Census, Washington, D.C.

LADKIN P. (1987). The logic of time representation. PhD, U. of California.

NADEAU D. & SEKINE S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, **30**(1), 3–26.

NATO (2005). *THE MILITARY INTELLIGENCE DATA EXCHANGE STANDARD - AIntP-3(B)*. Rapport interne.

NIST (2005). *The ACE 2005 (ACE05) Evaluation Plan*.

PUSTEJOVSKY J. *et al.* (2003). TimeML : Robust specification of event and temporal expressions in text. In *NDQA'03*, p. 28–34.

ROSARIO B. & HEARST M. A. (2005). Multi-way relation classification : application to protein-protein interactions. In *HLT-EMNLP'05*, p. 732–739.

SAVAL A., BOUZID M. & BRUNESSAUX S. (2009). A semantic extension for event modelisation. *21st IEEE ICTAI 2009*.

SERRANO L. *et al.* (2011). Extraction de connaissances pour le renseignement en sources ouvertes. In *EGC'2011 workshop SOS*.

SERRANO L. *et al.* (2012). Combinaison d'approches pour l'extraction automatique d'événements. In *TALN'2012*, France.

WINKLER W. E. (2006). *Overview of record linkage and current research directions*. Rapport interne, Bureau of the Census.

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