Vers un système de capitalisation des connaissances : extraction d'événements par combinaison de plusieurs approches

Laurie Serrano, Maroua Bouzid, Thierry Charnois, Bruno Grilhères

Vers un système de capitalisation des connaissances : extraction d'événements par combinaison de plusieurs approches.

Laurie Serrano*,** Maroua Bouzid*
Thierry Charnois*
Grilheres Bruno**

*GREYC, Université de Caen Basse-Normandie
prenom.nom@unicaen.fr, http://www.greyc.fr
**IPCC, Cassidian
prenom.nom@cassidian.com, http://weblab-project.org

Résumé. Face à l'augmentation vertigineuse de l'information disponible librement (en particulier sur le Web), repérer efficacement les informations qui sont susceptibles de nous intéresser (dans un contexte professionnel ou personnel) s'avère une tâche longue et complexe. En réponse à cela, l'équipe IPCC [1] développe le WebLab [2], une plateforme d'intégration de différents services de « media mining » [3] pour la découverte de connaissances et l'aide à la décision. Dans cet article, nous présentons notre système d'extraction automatique d'événements pour le ROSO [4] fondé sur la combinaison de plusieurs approches actuelles en extraction d'information. Nous proposons, dans un premier temps, une modélisation du domaine et plus particulièrement des événements. Puis, nous décrivons de façon détaillée notre approche et les différentes techniques utilisées. Enfin, nous concluons en résumant l'avancée de nos travaux et les perspectives envisagées.

1 Introduction

Nos travaux se placent dans le cadre du WebLab, plateforme open source dédiée à l'intégration d'outils de « media mining » et exploitant les technologies du Web sémantique (Giroux et al. (2008), Brunessaux et al. (2011)). Cet article présente des recherches en cours visant à exploiter efficacement la masse croissante d'informations disponibles en sources ouvertes afin d'en extraire un ensemble de connaissances pertinentes. Il s'agit, plus précisément, de proposer un système de capitalisation des connaissances visant à faciliter et réduire le travail des opérationnels dans le cadre de la veille économique et stratégique et du ROSO. Pour cela, nos travaux s'organisent selon trois axes de recherche :

1. Information Processing, Control and Cognition
2. `http://weblab-project.org/`
3. Fouille de documents multimedia
4. Renseignement d'Origine Sources Ouvertes

– extraction d'information ;
– capitalisation et gestion des connaissances ;
– interaction homme-machine.

La figure 1 donne un aperçu du fonctionnement général du système de capitalisation des connaissances que nous envisageons. Pour résumer, un ensemble de documents est traité par le système d'extraction d'information, puis les informations extraites sont stockées sous la forme de triplets RDF [5] dans une base de connaissances. Celle-ci est régie par notre ontologie de domaine (cf. partie 2) et couplée à un moteur d'inférence permettant la découverte de nouvelles connaissances. Chaque événement est ensuite présenté à l'utilisateur sous forme d'une fiche de connaissances que celui-ci pourra compléter et modifier en fonction de son propre savoir (cf. partie 4). Enfin, nous envisageons d'utiliser les différentes actions de l'utilisateur afin d'améliorer les performances futures du système.

Cet article est centré sur notre modèle d'extraction automatique d'événements fondé sur la combinaison de plusieurs approches actuelles en extraction d'information. Ici, nous souhaitons extraire un ensemble d'événements d'intérêt (préalablement définis dans une ontologie de domaine) à partir de dépêches de presse (AFP par exemple) en anglais et en français. Nous proposons, dans un premier temps, une modélisation du domaine et plus particulièrement des événements. Puis, nous décrivons de façon détaillée notre approche et les différentes techniques utilisées. Enfin, nous concluons en résumant l'avancée de nos travaux et les perspectives envisagées.

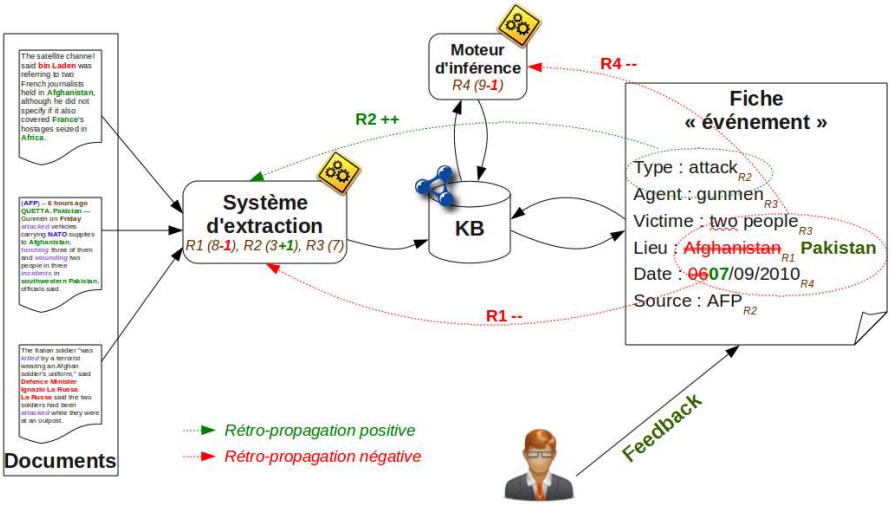

FIG. 1 – *Système de capitalisation des connaissances*

5. Resource Description Framework, `http://www.w3.org/RDF/`

2 Modélisation des connaissances

Afin de définir avec précision les informations pertinentes dans ce cadre d'application, celles-ci ont été modélisées sous la forme d'une ontologie de domaine.

2.1 WOOKIE : Weblab Ontology for Open sources Knowledge and Intelligence Exploitation

WOOKIE constitue l'ontologie de domaine qui sert de guide à notre système de capitalisation. Celle-ci est implémentée au format OWL [6] et suit les recommandations du W3C [7] concernant la représentation de la connaissance dans le cadre du Web Sémantique. WOOKIE a été élaborée suite à un état de l'art approfondi du domaine et des ontologies existant à l'heure actuelle. Nous avons examiné à la fois des ontologies dites de « haut niveau » mais également des modélisations spécifiques au ROSO telles que celles décrites par les standards OTAN. Suite à cet état de l'art, nous avons choisi de créer notre propre ontologie afin de répondre précisément aux besoins de notre application, en particulier pour la modélisation des événements. Toutefois, nous exploitons les travaux déjà réalisés par la création de liens (imports d'ontologies et équivalences de classes) entre WOOKIE et d'autres ontologies telles que OWL-Time [8] pour la représentation temporelle, Geonames [9] pour la représentation spatiale, FOAF [10], mais aussi The Event Ontology [11] et SEM [12] concernant les événements. WOOKIE est centrée sur le « pentagramme du renseignement » représentant les cinq entités centrales du ROSO (les événements, les personnes, les organisations, les lieux et les équipements) ainsi qu'un ensemble de liens entre ces entités (Serrano et al. (2011) pour plus de détails). Nous avons accordé une attention particulière à la modélisation de la classe « événement » compte tenu de l'importance de ces entités pour toute activité de veille (cf. partie 2.2). Précisons enfin que, même si WOOKIE est à l'heure actuelle utilisée dans le cadre du ROSO, nos choix de modélisation permettent d'éventuelles extensions à d'autres domaines d'application (intelligence économique, etc.).

2.2 Définition et modélisation des événements

L'événement étant l'objet central de notre système de capitalisation il est nécessaire de définir plus précisément ce concept. Considéré comme une entité aux propriétés bien spécifiques, l'événement a initialement été étudié par des philosophes (Davidson (2001)) puis par des linguistes (Van De Velde (2006)).

Nous avons retenu certains travaux tels que ceux de Krieg-Planque (2009). L'auteur donne une définition simple de l'événement mais qui nous paraît adaptée : « un événement est une occurrence perçue comme signifiante dans un certain cadre ». Ici, le terme « occurrence » met l'accent sur la notion de temporalité qui fait partie intégrante de ce concept. Cet aspect temporel de l'événement est au coeur de l'approche TimeML (Pustejovsky et al. (2003)), l'un des deux

6. Ontology Web Language, `http://www.w3.org/TR/owl-features/`
7. World Wide Web Consortium, `http://www.w3.org/`
8. `http://www.w3.org/TR/owl-time/`
9. `http://www.geonames.org/`
10. Friend Of A Friend, `http://www.foaf-project.org/`
11. `http://motools.sourceforge.net/event/event.html`
12. Simple Event Model, `http://www.few.vu.nl/~wrvhage/#research`

principaux courants dans le domaine de l'extraction automatique des événements. Le modèle utilisé dans le cadre des campagnes d'évaluation ACE (NIST (2005)) diffère de cette approche en définissant l'événement comme une structure complexe impliquant plusieurs arguments. Le « cadre » selon Krieg-Planque réfère à « un système d'attentes donné » qui « détermine le fait que l'occurrence acquiert (ou non) [...] sa remarquabilité [...] et, par conséquent, est promue (ou non) au rang d'événement. ». Dans nos travaux, ce cadre est défini par l'ontologie de domaine WOOKIE et plus précisément par la spécification de la classe « Event » à travers ses différentes sous-classes et propriétés. D'autre part, Neveu et Quéré (1996) s'attachent à décrire plus précisément la sémantique portée par les événements. Ils soulignent que l'interprétation d'un événement est étroitement liée au contenu sémantique des termes utilisés pour nommer cet événement. Pour plus de clarté nous reprendrons le terme de « nom d'événement » proposé par Krieg-Planque (2009). Ces « noms d'événement » transposent en langage naturel la « propriété sémantique » des événements mentionnée par Saval et al. (2009). De plus, cette description de l'événement est au centre d'un phénomène plus large, que Ricœur (1983) nomme « mise en intrigue », visant à organiser, selon le cadre mentionné plus haut, un ensemble d'éléments circonstants ou participants de l'événement.

En considérant ces différents travaux, nous présentons ci-après la définition d'un événement dans le cadre de nos recherches. Nous prenons pour point de départ la définition de Krieg-Planque citée précédemment. Toutefois, celle-ci étant très théorique, il convient d'expliciter comment un événement est exprimé au sein des dépêches de presse et comment est modélisé ce concept au sein de notre ontologie de domaine. Après observation de plusieurs dépêches, celles-ci semblent rapporter un événement principal, celui-ci étant le plus souvent résumé dans le titre et explicité tout au long de l'article (parfois en faisant référence à d'autres événements secondaires). En examinant de plus près cette description de l'événement tout au long de la dépêche, un certain nombre de « sous-événements » contribuent à la « mise en intrigue » mentionnée auparavant. Ces « sous-événements » correspondent à une association entre un « nom d'événement » et une ou plusieurs entités d'intérêt (une date, un lieu et des participants à l'événement). Dans le cadre de nos travaux, notre but est d'extraire ces « sous-événements » d'intérêt pour ensuite fusionner automatiquement ceux qui réfèrent dans la réalité à un seul et même événement. Chaque événement résultant de cette fusion sera présenté à l'utilisateur sous la forme d'une fiche de connaissances. Afin de proposer une représentation 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 d'un événement correspond aux différents types d'événement définis dans notre ontologie de domaine (les sous-classes de la classe « Event »), la composante temporelle constitue la date ou période d'occurrence de l'événement et l'entité spatiale équivaut au lieu d'occurrence de l'événement. 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 dorénavant $E\langle I, SP, S, A\rangle$ où A est un ensemble de participants jouant un ou plusieurs rôle(s). Un participant est noté P_i où $0 < i < n$ et un rôle est noté r_j où $0 < j < k$. 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. La figure 2 illustre notre modélisation de l'événement dans l'ontologie de domaine (un exemple d'événement est proposé en vert).

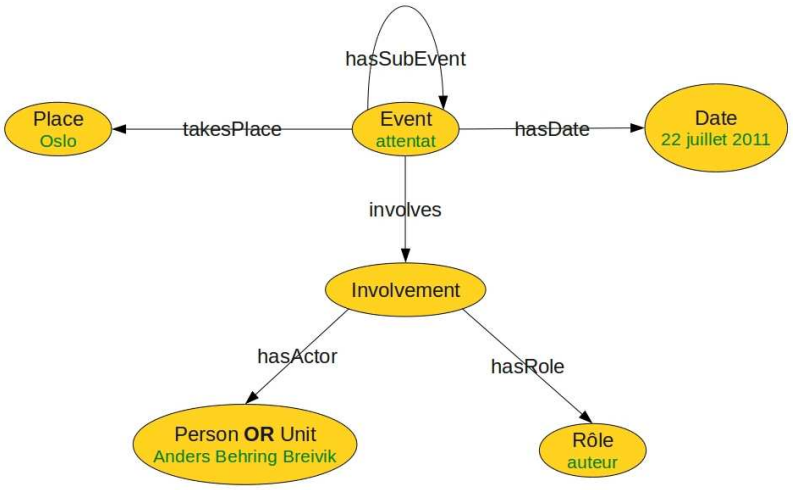

FIG. 2 – *Modélisation d'un événement dans WOOKIE*

3 Extraction d'information

3.1 Tour d'horizon

L'extraction d'information est une discipline assez récente qui consiste en une analyse partielle d'un texte afin d'en extraire des informations spécifiques. Celles-ci permettent de construire une représentation structurée (bases de données, fiches, tableaux) d'un document à l'origine non-structuré. Cela en fait une approche guidée par le but de l'application dans laquelle elle s'intègre, dépendance qui reste, à l'heure actuelle, une limite majeure des systèmes d'extraction. Les tâches les plus communes en extraction d'information restent la reconnaissance d'entités nommées (Nadeau et Sekine (2007)), de relations entre entités et d'événements (Hobbs et Riloff (2010)).

La reconnaissance d'entités nommées a été et est encore beaucoup étudiée car celles-ci constituent des éléments indispensables pour l'extraction d'entités plus complexes et reste une tâche difficile. Ces entités correspondent de façon générale aux noms de personne, organisation, lieu, mais aussi aux dates, unités monétaires, pourcentages, unités de mesure, etc. Par ailleurs, les différents outils d'extraction d'information s'attachent à extraire les relations entre entités et les événements. Les relations correspondent aux liens existant entre différentes entités repérées dans un texte : il peut s'agir par exemple de détecter les relations entre une personne et une organisation (appartenance, direction, etc.) ou encore d'extraire les attributs d'une personne (date de naissance, courrier électronique, adresse, etc.). Enfin, une dernière tâche est l'extraction d'événements, particulièrement utile dans les activités de veille économique et stratégique (Naughton et al. (2006)). Celle-ci peut-être conçue comme une forme

particulière d'extraction de relations où un « nom d'événement » est relié avec une date, un lieu et des participants (cf. partie 2.2).

Les dix dernières années ont vu apparaître un intérêt grandissant pour ce domaine avec notamment la création de campagnes d'évaluation telles que ACE [13], MUC [14], ESTER [15], CONLL [16], TAC [17], etc. Deux approches principales émergent alors : l'extraction basée sur des techniques linguistiques d'un côté et les systèmes statistiques à base d'apprentissage de l'autre. Celles-ci se basent, de façon commune, sur des pré-traitements linguistiques « classiques » comme la « tokenization » (découpage en mots), la lemmatisation (attribution de la forme non-fléchie associée), l'analyse morphologique (structure et propriétés d'un mot) ou syntaxique (structure d'une phrase et relations entre éléments d'une phrase). La première approche exploite les avancées en TAL [18] et repose principalement sur l'utilisation de grammaires formelles construites par la main d'un expert-linguiste. Les pré-traitements cités plus haut servent de base à la construction de règles et patrons linguistiques qui définissent les contextes d'apparition de telle entité ou relation. Notons ici l'importance particulière accordée à l'analyse syntaxique (en constituants ou dépendance) dans le repérage et le typage des relations et des événements. La seconde approche utilise des techniques statistiques pour « apprendre » des régularités sur de larges corpus de textes où les entités-cibles ont été préalablement annotées. Ces méthodes d'apprentissage sont supervisées, non-supervisées ou semi-supervisées et exploitent des caractéristiques textuelles plus ou moins linguistiques. Parmi celles-ci nous pouvons citer les « modèles de Markov Caché » (HMM), les « Conditional Random Fields » (CRF), les « Support Vector Machine » (SVM), etc. (Ireson et Ciravegna (2005) pour un état de l'art approfondi). Par ailleurs, de plus en plus de recherches portent sur l'apprentissage de ressources linguistiques ou encore sur l'utilisation d'un apprentissage dit « semi-supervisé » visant à combiner des données étiquetées et non-étiquetées (Nadeau (2007), Hobbs et Riloff (2010)).

Un nouveau type d'approche tend à se généraliser : ce sont les méthodes hybrides. Les acteurs du domaine choisissent de combiner plusieurs techniques face aux limites des approches symboliques et statistiques. Tout d'abord, celles-ci s'avèrent dépendantes d'un domaine ou d'un genre de texte particulier et cela nécessite une constante ré-adaptation des modèles d'extraction. Les approches à base de règles linguistiques souffrent également d'un développement manuel coûteux et de la nécessité d'une expertise en linguistique pour pouvoir les modifier et les adapter. Pour tenter de résoudre cela, les experts se penchent actuellement vers des méthodes d'apprentissage automatique de patrons linguistiques (Charnois et al. (2009)). Pour finir, les approches statistiques nécessitent, lors de la phase d'apprentissage, une grande quantité de textes pré-annotés et cela constitue une réelle contrainte car ces données ne sont pas toujours disponibles. Des recherches sont menées dans le sens d'un apprentissage dit « semi-supervisé » visant à mêler des données étiquetées et non-étiquetées (Nadeau (2007)).

13. Automatic Content Extraction, http://www.itl.nist.gov/iad/mig/tests/ace/
14. Message Understanding Conference, http://www-nlpir.nist.gov/related_projects/muc/
15. Évaluation des Systèmes de Transcription Enrichie d'Émissions Radiophoniques, http://www.afcp-parole.org/camp_eval_systemes_transcription/
16. Conference on Computational Natural Language Learning, http://ifarm.nl/signll/conll/
17. Text Analysis Conference, http://www.nist.gov/tac/
18. Traitement Automatique des Langues

3.2 Modèle proposé

Dans la lignée de ces nouvelles approches, la combinaison de plusieurs techniques d'extraction nous parait être la solution la plus pertinente. En effet, élaborer un système composite permet de tirer le meilleur parti des différentes approches actuelles. Pour cela, nous avons choisi ici de combiner trois extracteurs :

- un extracteur symbolique, fondé sur des règles linguistiques écrites manuellement ;
- un extracteur statistique utilisant les CRF (Conditional Random Fields) ;
- un extracteur hybride basé sur l'apprentissage de motifs séquentiels fréquents.

3.2.1 Approche symbolique

Cette première approche a été développée grâce à la plateforme open source GATE [19]. Cet extracteur symbolique est composé d'un ensemble de règles de grammaire écrites manuellement et définissant les différents contextes d'apparition possible pour un type d'événement. Ces règles s'appuient sur un premier extracteur d'entités nommées que nous avons développé, puis sur une analyse syntaxique en dépendance (fournie par un module GATE) permettant d'établir des liens entre ces entités au niveau phrastique et de repérer les événements potentiellement pertinents pour notre domaine d'application. Une première évaluation en reconnaissance d'entités nommées a montré des résultats comparables à l'état de l'art avec de bonnes performances pour l'extraction des dates et des lieux mais aussi quelques faiblesses dans la délimitation des organisations ou des personnes. L'évaluation des événements extraits sera réalisée prochainement mais nous pouvons d'ores et déjà remarquer que les résultats de cet extracteur sont très fortement dépendants de l'exactitude de l'analyse syntaxique. Le fonctionnement et l'évaluation de ce système sont décrits plus en détails par Serrano et al. (2011).

3.2.2 Approche statistique

La seconde approche que nous avons choisie est basée sur l'utilisation des CRF [20], une technique statistique couramment utilisée en extraction d'information. Les CRF constituent un modèle probabiliste graphique discriminant et non orienté. Ce type de modèle et souvent utilisé pour annoter ou segmenter des séquences de données telles que des textes en langage naturel ou des séquences d'ADN. En traitement automatique des langues, les CRF trouvent des applications dans l'analyse syntaxique de surface (en constituants), la reconnaissance d'entités nommées et constituent une extension des Modèles de Markov Cachés (HMM). Dans le cadre de nos recherches, nous utilisons l'outil open source CRF++ [21] ainsi qu'un corpus annoté en entités nommées issu de la campagne d'évaluation TempEval [22]. L'apprentissage du modèle est réalisé en prenant en compte trois caractéristiques :

- la forme fléchie d'un mot ;
- sa catégorie grammaticale ;
- le type d'entité auquel il fait référence (le cas échéant).

19. General Architecture for Text Engineering, `http://gate.ac.uk/`
20. Conditional Random Fields
21. `http://crfpp.sourceforge.net/`
22. `http://timeml.org/site/timebank/timebank.html`

Cette partie de nos travaux est en cours et nous envisageons d'exploiter d'autres caractéristiques pour un apprentissage plus performant (forme lemmatisée, syntagmes, etc.). Les travaux existants ainsi que nos premières expérimentations ont montré que les résultats de ce type d'approche dépendent essentiellement de ces caractéristiques ainsi que de la quantité et de la qualité des données d'apprentissage. Nous sommes également en phase d'exploration des différentes techniques d'adaptation des CRF pour l'extraction de relations entre entités dans le cadre de l'extraction d'entités complexes telles que les événements.

3.2.3 Approche hybride

Enfin, nous nous intéressons à l'extraction d'événements par une technique hybride d'extraction de motifs séquentiels fréquents [23]. Ce type d'approche fait le lien entre les méthodes symboliques et statistiques en proposant d'apprendre automatiquement des patrons linguistiques.

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 »). Un certain nombre de paramètres peuvent être adaptés selon l'application visée : nature de la séquence et des items, nombre d'items, grain (mot, syllabe, paragraphe, etc.), support, etc. La fouille sur un ensemble d'items 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> [24], etc. De plus, contrairement aux différentes approches que nous venons de mentionner, l'apprentissage de MSF 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 encore une technologie peu disponible librement et aux performances inégales 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, Charnois et al. (2009) propose l'ajout de contraintes pour diminuer la quantité de motifs retournés.

Dans la lignée de ces travaux, nous utilisons l'outil CloSpan (Closed Sequential Pattern Mining Package) fourni dans le cadre du projet open source IlliMine [25]. Celui-ci s'avère très utilisé dans la communauté et 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 d'itemsets, fonctionnalité qui est rarement proposée par les outils existants. Nous sommes en train d'adapter la fouille de MSF avec CloSpan 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. Tout d'abord, nous avons pré-traité notre corpus grâce à l'outil TreeTagger (Schmid (1994)) afin d'obtenir un découpage en séquences (ici en phrases) ainsi que différents types d'items : mot (forme fléchie),

23. MSF par la suite
24. N pour la catégorie nom, PRP pour préposition
25. `http://illimine.cs.uiuc.edu/`

lemme, catégorie grammaticale. Nous utilisons également notre système symbolique pour obtenir un item « entité nommée » (cela est provisoire, nous envisageons de réaliser une fouille de MSF dédiée à cette tâche). Enfin, nous effectuons un repérage lexical des potentiels « noms d'événement » et de leur type (sous-classes d'événement dans l'ontologie). Comme prévu, le nombre de motifs retournés par CloSpan s'avère élevé, nous devons donc introduire un ensemble de contraintes spécifiques à notre application. Dans un premier temps, seuls les motifs contenant au minimum un « nom d'événement » et une entité (lieu, date, personne, organisation) sont conservés. Les motifs sont ensuite regroupés selon le type de ce « nom d'événement » afin de dégager les régularités spécifiques à chaque type d'événement.

3.3 Qualité des extractions

Afin de combiner ces trois approches pour un système d'extraction plus performant, il devient nécessaire d'évaluer la qualité des extractions obtenues. En effet, cette évaluation servira, tout d'abord, en amont de la base de connaissances pour fusionner les résultats des différents systèmes d'extraction mais également lors de la présentation des fiches de connaissances à l'utilisateur (cf partie 4). L'évaluation de l'information est une problématique largement rencontrée dans la communauté du traitement de l'information et ceci à différents niveaux. Notre étude de la littérature à ce sujet a révélé d'une part des recherches portant sur la modélisation et d'autre part sur des techniques d'estimation de cette qualité.

Traitant des aspects modélisation, nous avons retenu ceux de Laskey et Laskey (2008) définissant une ontologie de « haut niveau » dédiée à la représentation de l'incertitude et visant à faciliter les mécanismes de raisonnement prenant en compte cette incertitude. Par ailleurs, Dedek et al. (2008) propose une extension de cette ontologie dédiée à la représentation de la qualité dans le domaine de l'extraction d'information. Nous avons également noté que la modélisation de la qualité de l'information est étroitement liée, dans beaucoup de travaux actuels, à la notion de provenance de l'information. En effet, dans beaucoup d'applications évaluer la qualité d'une information implique de connaître d'où elle provient, c'est-à-dire sa source mais aussi les différents traitements qui ont été opérés depuis sa collecte jusqu'à sa présentation à l'utilisateur. Davide Ceolin (2010) propose OPM [26], un modèle commun pour tracer et échanger la provenance d'une information. Dans la lignée de ces travaux, Van Hage et al. (2011) propose une combinaison des ontologies SEM et OPM dans le but d'estimer la confiance d'un événement. Sur des aspects plus pratiques, il existe des extensions de la syntaxe RDF par des « graphes nommés » permettant de faire des assertions sur des graphes [27] et, par ce biais, de représenter des méta-informations telles que la confiance et la provenance. Enfin, nous pouvons citer un ensemble de travaux portant sur différentes techniques d'évaluation de la qualité en extraction d'information. Van Keulen et Habib (2011) évalue la qualité des extractions grâce à un ensemble de règles de connaissance définies dans une ontologie combinée à une base de données probabiliste. Soderland et al. (2004) présente un module d'évaluation (intégré au sein de l'extracteur d'événements KnowItAll) fondé sur l'utilisation du Web combiné à un calcul d'information mutuelle (Pointwise Mutual Information). Nous nous sommes également intéressés aux recherches de Besombes et Revault D'Allonnes (2008) qui constituent une bonne

26. Open Provenance Model, `http://openprovenance.org/`
27. `http://www.w3.org/2009/12/rdf-ws/papers/ws06/`

approche générale de cotation de l'information prenant en compte les différents types d'incertitude mentionnés plus haut.

Au vu de ces différents travaux, nous sommes actuellement en cours de réflexion pour proposer une modélisation et une estimation de la qualité des extractions au sein de notre système de capitalisation des connaissances. Tout d'abord, la combinaison de plusieurs extracteurs est une particularité que nous devons prendre en compte dans nos choix. En effet, nous avons pu constater qu'il n'est pas évident d'évaluer et de modéliser la qualité d'une extraction de la même manière pour les 3 approches d'extraction que nous utilisons. Premièrement, l'extraction statistique à base de CRF est souvent qualifiée de « boite noire » car le modèle d'extraction appris est difficilement accessible et modifiable. Toutefois, ce système donne nativement une probabilité à chaque extraction retournée qui peut constituer un premier indice de qualité de l'information. L'extracteur symbolique, quant à lui, n'estime pas par défaut la qualité de ses résultats mais les règles d'extraction sont accessibles. Cette observation vaut aussi pour l'apprentissage symbolique car celui-ci aboutit à la création de règles du même type. Il est également possible d'évaluer les extractions produites par cette approche grâce au nombre d'apparitions de chaque motif dans le corpus d'apprentissage (dit « support »). Enfin, de manière générale, ces trois approches peuvent être jugées manuellement ou automatiquement par une évaluation « a priori » telle que celles menées par les campagnes d'évaluation du domaine (calcul de F-mesure, précision, rappel et autres métriques). Ces premières observations font émerger plusieurs questions importantes auxquelles nous devrons répondre : peut-on concevoir une méthode d'évaluation de cette qualité commune aux 3 extracteurs ? Devons-nous estimer la qualité du système d'extraction en lui-même (des règles ou des modèles appris) ou plutôt la qualité des extractions produites ? Un dernier problème est le mode de représentation de cette qualité : il s'agira de choisir la modélisation la plus proche des besoins du système mais aussi la plus compatible avec les technologies utilisées au sein du WebLab.

4 Fiches de connaissances

Toutes les informations extraites et stockées dans la base de connaissances sont présentées à l'utilisateur sous la forme de fiches de connaissances. Nous nous focalisons, ici, sur la construction automatique de fiches contenant les informations connues à propos d'un événement. Ce mode de visualisation apparaît comme le plus proche des besoins actuels des opérationnels du ROSO. Toutefois, il faut préciser que la fiche de connaissances est un moyen parmi d'autres de visualiser l'information et n'impacte pas la représentation interne de la base de connaissances. Pour construire ces fiches automatiquement, il est nécessaire de définir la nature des informations présentées à l'utilisateur mais aussi les différentes interactions possibles avec le système de capitalisation des connaissances.

Tout d'abord, la fiche comporte plusieurs champs : une description générale (type, nom, alias), la situation dans laquelle cet événement est survenu (lieu, date/période), les différents participants impliqués dans l'événement et enfin un certain nombre de liens avec d'autres événements découverts grâce au moteur d'inférence. Face à une fiche, l'utilisateur a la possibilité de modifier et/ou valider ces différents champs. Il peut alors effectuer plusieurs types d'action :

- valider un champ ou la fiche complète ;
- ajouter un nouveau champ ;
- corriger ou supprimer un champ.

Il faudra bien sûr mettre à jour automatiquement la base de connaissances selon ces modifications. Par ailleurs, l'utilisateur sera guidé dans son activité de veille par le système d'évaluation de la fiabilité des informations abordé plus haut (cf. partie 3.3).

5 Conclusions et perspectives

Nous avons présenté ici une approche d'extraction d'information fondée sur la combinaison de plusieurs extracteurs dans le but d'obtenir de meilleures performances. L'évaluation de notre extracteur symbolique ayant montré des résultats comparables à l'état de l'art mais encore imparfaits, nous attendons de meilleures performances grâce à l'apport des systèmes statistique et hybride. Notre approche se distingue par cette combinaison mais également par la phase de fusion des extractions suggérées par ces différents systèmes qui constitue une phase essentielle dans la capitalisation des connaissances. Nous explorons actuellement un certain nombre de travaux menés en fusion d'informations dans le but de les adapter à la fusion d'informations textuelles. De plus, nous voulons mettre l'accent sur l'importance de l'évaluation de la qualité des extractions pour aider à cette étape de fusion mais également pour guider l'utilisateur dans son travail de validation des fiches. D'autre part, le second axe de nos recherches porte sur les problématiques de capitalisation des connaissances à travers les mécanismes de raisonnement et d'inférence. Nous nous attacherons à définir des règles d'inférence permettant à la fois de regrouper différents « sous-événements » en un seul événement d'intérêt mais également de découvrir des liens entre événements. Il s'agira non seulement de raisonnement spatio-temporel mais aussi d'inférence tenant compte des relations entre un événement et ses participants, tous deux encadrés par notre représentation ontologique du domaine. Par ailleurs, le retour de l'utilisateur étant à l'heure actuelle indispensable face aux limites des technologies employées, nous exploiterons celui-ci afin de réévaluer l'estimation de confiance des extracteurs. Enfin, nous définirons une méthodologie d'évaluation de nos travaux qui comprendra, d'une part, une évaluation automatique du système d'extraction et, d'autre part, une évaluation globale du système de capitalisation par un ensemble d'utilisateurs.

Références

Agrawal, R., T. Imieliński, et A. Swami (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, SIGMOD '93, New York, pp. 207–216. ACM.

Besombes, J. et A. Revault D'Allonnes (2008). An extension of STANAG2022 for information scoring. In *Fusion 2008 - 11th International Conference on Information Fusion*, Allemagne, pp. 1635–1641.

Brunessaux, S., S. Cantarell, A. Giraudel, G. Patrick, et G. Bruno (2011). Herisson : a weblab-based platform for assessment and experimentation of information processing technologies. In *ICSSEA*.

Cellier, P. et T. Charnois (2010). Fouille de données séquentielle d'itemsets pour l'apprentissage de patrons linguistiques. In *Traitement Automatique des Langues Naturelles (short paper)*.

Charnois, T., M. Plantevit, C. Rigotti, et B. Cremilleux (2009). Fouille de données séquentielles pour l'extraction d'information dans les textes. *Revue Traitement Automatique des Langues (TAL) 50*(3), 59–87.

Davide Ceolin, P. G. (2010). Calculating the trust of event descriptions using provenance.

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

Dedek, J., A. Eckhardt, L. Galambos, et P. Vojtás (2008). Discussion on uncertainty ontology for annotation and reasoning. In *URSW*.

Giroux, P., S. Brunessaux, S. Brunessaux, J. Doucy, G. Dupont, B. Grilheres, Y. Mombrun, et A. Saval (2008). Weblab : An integration infrastructure to ease the development of multimedia processing applications. *ICSSEA*.

Hobbs, J. R. et E. Riloff (2010). Information extraction. In *Handbook of Natural Language Processing, Second Edition*. Boca Raton, FL : CRC Press, Taylor and Francis Group.

Ireson, N. et F. Ciravegna (2005). Pascal challenge the evaluation of machine learning for information extraction. In *Proceedings of Dagstuhl Seminar Machine Learning for the Semantic Web*.

Krieg-Planque, A. (2009). *A propos des noms propres d'événement*, Volume 11, pp. 77–90. Les carnets du Cediscor.

Laskey, K. J. et K. B. Laskey (2008). Uncertainty Reasoning for the World Wide Web : Report on the URW3-XG Incubator Group. In *URSW'08*.

Nadeau, D. (2007). *Semi-Supervised Named Entity Recognition : Learning to Recognize 100 Entity Types with Little Supervision*. Ph. D. thesis.

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

Naughton, M., N. Kushmerick, et J. Carthy (2006). Event extraction from heterogeneous news sources. In *Proc. Workshop Event Extraction and Synthesis*. American Nat. Conf. Artificial Intelligence.

Neveu, E. et L. Quéré (1996). *Le temps de l'événement I*, Chapter Présentation, pp. 7–21. Number 75 in Réseaux. CNET.

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

Pustejovsky, J., J. M. Castaño, R. Ingria, R. Sauri, R. J. Gaizauskas, A. Setzer, G. Katz, et R. D. R. (2003). TimeML : Robust specification of event and temporal expressions in text. In *New Directions in Question Answering'03*, pp. 28–34.

Ricœur, P. (1983). *Temps et récit 1. L´intrigue et le récit historique*, Volume 227 of *Points : Essais*. Paris : Ed. du Seuil.

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

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Serrano, L., B. Grilheres, M. Bouzid, et T. Charnois (2011). Extraction de connaissances pour le renseignement en sources ouvertes. In *Atelier Sources Ouvertes et Services (SOS'2011) en conjonction avec la conférence internationale francophone EGC'2011*.

Soderland, S., O. Etzioni, T. Shaked, et D. S. Weld (2004). The use of web-based statistics to validate information extraction. In *Proceedings of AAAI 2004 Workshop on Adaptive Text Extraction and Mining (ATEM'04)*.

Van De Velde, D. (2006). *Grammaire des événements*. Presses Universitaires du Septentrion.

Van Hage, W. R., V. Malaisé, R. H. Segers, L. Hollink, et G. Schreiber (2011). Design and use of the simple event model (sem). *Web Semantics : Science, Services and Agents on the World Wide Web 9*(2).

Van Keulen, M. et M. B. Habib (2011). Handling uncertainty in information extraction. In *URSW*, Volume 778 of *CEUR Workshop Proceedings*, pp. 109–112.

Summary

Nowadays staggering increasing of the information available (specially on the Web) makes the discovery of relevant information (professionnally or personnaly speaking) more and more complex and time-consuming. Tackling this issue, the IPCC team develops the WebLab, a media mining platform aiming at integrating various tools to enhance knowledge discovery and decision making. In this paper, we present an automatic system to extract events dedicated to open source intelligence and based on the combination of multiple information extraction approaches. Firstly, we propose a quick state-of-the-art of the main scientific axes our researches tackle. Then, we describe more precisely our model and all the techniques we use. To conclude, we summarize what we have done so far and the future work we plan to do.