Vers une capitalisation des connaissances orientée utilisateur : extraction et structuration automatiques de l'information issue de sources ouvertes

Laurie Serrano

Université de Caen Basse-Normandie

École doctorale SIMEM

Thèse de doctorat

présentée et soutenue le : *24/01/2014*

par

Laurie Serrano

pour obtenir le

Doctorat de l'Université de Caen Basse-Normandie

Spécialité : Informatique et applications

Vers une capitalisation des connaissances orientée utilisateur

Extraction et structuration automatiques de l'information issue de sources ouvertes

Directrice de thèse : *Maroua Bouzid*

Jury

Laurence Cholvy	Directrice de recherche	DTIM, ONERA	(Rapporteur)
Thierry Poibeau	Directeur de recherche	LaTTiCe, ENS	(Rapporteur)
Fatiha Saïs	Maître de conférences	LRI, Univ. Paris 11	(Examinatrice)
Gaël Dias	Professeur des universités	GREYC, Univ. de Caen	(Examinateur)
Stephan Brunessaux	Senior expert	Cassidian, EADS	(Co-directeur de thèse)
Thierry Charnois	Professeur des universités	LIPN, Univ. Paris 13	(Co-encadrant de thèse)
Maroua Bouzid	Professeur des universités	GREYC, Univ. de Caen	(Directrice de thèse)

Mis en page avec la classe thloria.

Résumé

Face à l'augmentation vertigineuse des informations disponibles librement (notamment sur le Web), repérer efficacement celles qui présentent un intérêt s'avère une tâche longue et complexe. Les analystes du renseignement d'origine sources ouvertes sont particulièrement concernés par ce phénomène. En effet, ceux-ci recueillent manuellement une grande partie des informations d'intérêt afin de créer des fiches de connaissance résumant le savoir acquis à propos d'une entité. Dans ce contexte, cette thèse a pour objectif de faciliter et réduire le travail des acteurs du renseignement et de la veille. Nos recherches s'articulent autour de trois axes : la modélisation de l'information, l'extraction d'information et la capitalisation des connaissances. Nous avons réalisé un état de l'art de ces différentes problématiques afin d'élaborer un système global de capitalisation des connaissances. Notre première contribution est une ontologie dédiée à la représentation des connaissances spécifiques au renseignement et pour laquelle nous avons défini et modélisé la notion d'événement dans ce domaine. Par ailleurs, nous avons élaboré et évalué un système d'extraction d'événements fondé sur deux approches actuelles en extraction d'information : une première méthode symbolique et une seconde basée sur la découverte de motifs séquentiels fréquents. Enfin, nous avons proposé un processus d'agrégation sémantique des événements afin d'améliorer la qualité des fiches d'événements obtenues et d'assurer le passage du texte à la connaissance. Celui-ci est fondé sur une similarité multidimensionnelle entre événements, exprimée par une échelle qualitative définie selon les besoins des utilisateurs.

Mots-clés: Gestion des connaissances, exploration de données, représentation des connaissances, renseignement d'origine sources ouvertes, ontologies (informatique), Web sémantique.

Abstract

Due to the considerable increase of freely available data (especially on the Web), the discovery of relevant information from textual content is a critical challenge. Open Source Intelligence (OSINT) specialists are particularly concerned by this phenomenon as they try to mine large amounts of heterogeneous information to acquire actionable intelligence. This collection process is still largely done by hand in order to build knowledge sheets summarizing all the knowledge acquired about a specific entity. Given this context, the main goal of this thesis work is to reduce and facilitate the daily work of intelligence analysts. For this sake, our researches revolve around three main axis : knowledge modeling, text mining and knowledge gathering. We explored the literature related to these different domains to develop a global knowledge gathering system. Our first contribution is the building of a domain ontology dedicated to knowledge representation for OSINT purposes and that comprises a specific definition and modeling of the event concept for this domain. Secondly, we have developed and evaluated an event recognition system which is based on two different extraction approaches : the first one is based on hand-crafted rules and the second one on a frequent pattern learning technique. As our third contribution, we proposed a semantic aggregation process as a necessary post-processing step to enhance the quality of the events extracted and to convert extraction results into actionable knowledge. This is achieved by means of multiple similarity measures between events, expressed according a qualitative scale which has been designed following our final users' needs.

Keywords: Knowledge management, data mining, knowledge representation (information theory), open source intelligence, ontologies (information retrieval), Semantic Web.

Remerciements

Cette thèse ayant été réalisée dans le cadre d'une convention industrielle, je tiens tout d'abord à remercier les personnes de mon entreprise et de mon laboratoire qui en sont à l'origine : notamment Stephan Brunessaux et Bruno Grilheres pour m'avoir recrutée et encadrée en stage puis proposé cette collaboration, mais aussi mes encadrants académiques, Maroua Bouzid et Thierry Charnois. Je vous suis à tous extrêmement reconnaissante pour le temps que vous m'avez consacré du début à la fin de cette thèse ainsi que pour toutes nos discussions enrichissantes qui ont permis à mes travaux d'avancer.

Je tiens par ailleurs à remercier l'ensemble des membres du jury : Laurence Cholvy et Thierry Poibeau pour avoir accepté de rapporter mon travail ainsi que Fatiha Saïs et Gaël Dias pour leur participation au jury de soutenance.

Mes remerciements les plus sincères vont également aux deux équipes au sein desquelles j'ai été intégrée pendant ces trois ans. Je remercie tous les membres de l'équipe MAD pour leur accueil et leurs conseils notamment lors des groupes de travail. Mais aussi et bien sûr tous les membres de l'équipe IPCC : vous m'avez accueillie les bras ouverts et je garderai une place pour vous tous dans ma petite tête, je ne pouvais pas rêver mieux comme équipe pour une première expérience professionnelle. Je te remercie encore Stephan de m'avoir recrutée, Bruno pour ton encadrement justement dosé qui m'a guidé tout en favorisant mon autonomie, Khaled mon cher collègue de bureau pour tes conseils mais aussi pour nos rires même dans les moments de *rush*. Yann, Arnaud, Emilien, Amandine pour votre aide précieuse quand je venais vous embêter avec mes questions et tous les autres bien sûr qui m'ont spontanément apporté leur aide et soutien. Fred (je ne t'oublie pas non non), notre "chef de centre" farceur et toujours là pour entretenir cette bonne humeur qui caractérise notre équipe, je te remercie pour toutes nos rigolades. Toi aussi, Dafni, ma grecque préférée, je te dois énormément, professionnellement et personnellement, tu as su me comprendre et me conseiller à tout moment et je ne l'oublierai pas. Véro, ma tata, mon amie, tu as été là pour moi depuis le tout premier jour quand tu es venue me chercher à la gare et que nous avons tout de suite sympathisé. Je te remercie pour tout ce que tu as fait pour moi et tout ce que tu continues d'apporter à cette équipe avec ton grand cœur.

Je pense aussi à mes amis qui ont été présents pendant cette aventure, tout proche ou à distance, ponctuellement ou en continu, dans les moments difficiles ou pour le plaisir, mais tous toujours là. Milie, Manon, Laure, Lucile, Rosario, Chacha, Marlou, je vous remercie mes chers Talistes/Taliens pour votre soutien, votre écoute, votre folie, nos retrouvailles, nos fiestas et bien d'autres moments passés avec vous. Un gros merci également à mes amis et colocs rouennais, Romain, Etienne, mon Aldricou, Nico, Juline, Manuella, Camille et bien d'autres, vous avez su me changer les idées durant nos traditionnelles soirées à l'Oka, à la coloc, nos escapades au ski, à Brighton et ailleurs. Je suis fière de vous avoir rencontrés et je ne vous oublierai pas. Une pensée pour toi aussi ma Lady, toi qui m'a soutenue dans l'un des moments les plus difficiles de cette thèse. Je vous remercie Julien, Xavier, Mariya et Laura, mes chers amis de Caen, avec vous j'ai pu partager mes petits soucis de thésarde et passer de très agréables soirées caennaises. *Last but not least*, vous mes Trouyiens adorés, mes amis de toujours, éparpillés en France et ailleurs, Bolou, Maelion, Solenou, Loicou, Tildou, Emilie et Quitterie, sans même vous en rendre compte, vous m'avez aidée pendant cette thèse oui oui, car cette thèse n'est qu'une petite partie d'une aventure bien plus enrichissante...

Enfin, je veux remercier mes parents, *los de qui cau*, qui, *drets sus la tèrra*, m'ont transmis des valeurs chères à mes yeux et m'ont poussée à aller au bout de ce que j'entreprends et de ce que j'aime.

Table des matières

Table des figures

Liste des tableaux

Introduction

Sommaire

Le savoir occupe aujourd'hui et depuis toujours une place centrale dans notre société, il est au cœur de toute activité humaine. Épanouissement intellectuel pour certains et capital pour d'autres, la connaissance est considérée comme une richesse pouvant servir un large panel d'objectifs. Que ce soit à des fins personnelles ou professionnelles, la principale visée de l'acquisition du savoir quel qu'il soit est, sans aucun doute, de mieux appréhender et de comprendre notre environnement. Dans des situations variées et en constante mutation, la capacité à observer et à analyser ce qui nous entoure (objets, personnes, situations, relations, faits, etc.) est un préalable fondamental à tout processus de prise de décision.

L'essor d'Internet et des nouvelles technologies de l'information a récemment déstabilisé les principaux mécanismes traditionnels de gestion de la connaissance. Passé d'un ensemble restreint et structuré de silos à la Toile, le savoir est de plus en plus accessible à tous et partout. Ce changement provient principalement de la démocratisation des moyens de communication et de publication de l'information. Deux problématiques principales émergent alors :

– Comment faire face à cette nouvelle masse disponible qui constitue une mine d'or mais peut également s'avérer néfaste à notre acquisition de connaissance ?
– Quels moyens mettre en place pour extraire un savoir homogène à partir de contenus de plus en plus diversifiés sur le fond et sur la forme ?

Celles-ci occupent une place centrale dans divers domaines pour lesquels l'acquisition du savoir est stratégique : le Renseignement d'Origine Sources Ouvertes (ROSO) est l'un de ces domaines.

1 Contexte

En guise d'introduction de nos travaux, nous proposons tout d'abord un rapide tour d'horizon du contexte de cette étude, réalisée dans un cadre à la fois académique et applicatif. Nous parlerons, dans un premier temps, du Renseignement d'Origine Sources Ouvertes constituant le fil directeur de nos recherches en termes de besoins opérationnels. Puis, nous introduirons un champ de recherche visant à répondre à ces besoins et dans lequel se situe plus précisément notre sujet de recherche, à savoir la fouille de documents multimédia, plus communément désignée par le terme de *Media Mining*.

1.1 Renseignement d'Origine Sources Ouvertes

Le Renseignement d'Origine Sources Ouvertes (dit ROSO) désigne toute activité de recueil et d'analyse de l'information disponible publiquement et légalement (presse écrite, blogs, sites internet, radio, télévision, etc.) [Best and Cumming, 2007]. Initialement définie dans le domaine de la défense, cette activité est aujourd'hui menée plus largement à des fins stratégiques et économiques sous le nom de veille à partir de sources ouvertes.

En effet, les Sources Ouvertes (SO) sont très prolifiques et peuvent, par exemple, fournir les données nécessaires pour analyser la situation d'un pays : caractéristiques géopolitiques et sociales, différents acteurs économiques, politiques, militaires, terroristes ou criminels, etc. Lors d'une crise, l'analyse systématique des médias nationaux et internationaux peut permettre, par exemple, de produire automatiquement une synthèse qui facilitera les prises de décision. Des processus de veille peuvent également être mis en œuvre pour effectuer une recherche pro-active d'informations liées à l'environnement, aux opportunités ou aux menaces qui constituent des signaux faibles et qui doivent faire l'objet d'une écoute anticipative.

Les objectifs premiers du ROSO sont les suivants :

- Savoir : s'informer sur les intentions d'un acteur intérieur ou extérieur, comprendre cet acteur et la situation ;
- Prévoir : anticiper les évolutions, prévenir les menaces, influencer les situations.

Le cycle du renseignement a été défini pour répondre à ces deux objectifs et est constitué de 5 étapes :

1. Orientation : il s'agit de définir le besoin en renseignement et spécifier les indicateurs permettant de valider la réussite de l'action de renseignement,
2. Planification : elle consiste à trouver les sources et les gisements d'information d'intérêt et à définir le besoin de veille,
3. Recherche : elle vise à réaliser l'acquisition et le pré-traitement des données à partir des gisements précédemment définis,
4. Exploitation : il s'agit d'analyser le contenu des données, de les filtrer pour en faire émerger de la connaissance,
5. Diffusion : elle consiste à faire la synthèse et à remonter l'information utile vers le décideur.

Le ROSO est aujourd'hui devenu un processus complexe à mettre en œuvre. En effet, avec la croissance du Web 2.0 et la multiplication des gisements d'information, les spécialistes du renseignement

et les veilleurs se trouvent confrontés, d'une part, à une masse de données toujours plus importante et, d'autre part, à une diversité croissante des formats et structures. Face à ce nouveau phénomène, des systèmes d'information plus performants sont désormais nécessaires pour accéder et traiter cette masse d'information. Ces systèmes sont notamment indispensables pour dépasser les limites des moteurs de recherche "grand public" et proposer une collecte ciblée et précise de l'information à la fois dans le Web public, mais aussi dans le Web profond. Ils doivent également être capables de détecter les informations pouvant constituer des signes précoces de changement ou de menace. Dès lors, le développement et l'utilisation de tels outils deviennent des tâches de plus en plus complexes.

L'essor des nouveaux moyens de communication favorise l'émergence des sources d'information mettant à disposition des informations aux caractéristiques particulières. La prise en compte de celles-ci est primordiale pour obtenir un système adapté à la fois aux besoins des analystes du renseignement mais également aux informations que ceux-ci sont amenés à traiter. Ainsi, les principales problématiques rencontrées lors du traitement des informations issues de sources ouvertes sont les suivantes :

- L'hétérogénéité des formats et structures : les informations disponibles sont proposées dans des formats variés (pages HTML, flux RSS, réseaux sociaux, wiki, blogs, forums, etc.) et ne sont pas toujours structurées. Le traitement de ces ressources implique l'utilisation d'un ensemble varié d'outils dont l'interopérabilité n'est pas assurée. Pour atteindre ses objectifs, le veilleur doit se former à leur utilisation combinée, ce qui augmente encore la complexité de son travail.
- Le multilinguisme : avec la démocratisation de l'accès à Internet, le nombre de langues employées sur la Toile a fortement augmenté ces dernières années, ce qui pose des problèmes d'intercompréhension. Les outils de traduction automatique deviennent donc clés afin de donner un accès à ces contenus dans des langues non maitrisées par l'analyste.
- La quantité d'information à traiter : la quantité et le volume des informations mises à disposition aujourd'hui en sources ouvertes, notamment avec la croissance des contenus audio et vidéo en ligne, sont tels qu'il devient impossible de collecter manuellement toutes ces informations. En effet, la collecte de ces grands volumes nécessite des connexions réseaux rapides, du temps, ainsi qu'un espace de stockage important, dont on ne dispose généralement pas. De plus, face à cette quantité d'informations disponibles, le veilleur se trouve submergé et il devient impossible pour lui de traiter efficacement ces nouvelles données, et de discerner clairement les informations pertinentes pour sa tâche.
- La qualité et l'interprétation des informations : les informations disponibles en sources ouvertes peuvent être peu fiables, contradictoires, dispersées, et il s'avère souvent difficile, voire impossible de savoir quel crédit leur accorder au premier abord. Il convient alors de recouper et d'analyser ces informations pour leur donner un sens et une valeur, ce qui reste une étape manuelle et donc coûteuse, à la fois en termes de temps et de moyens. Comment rassembler et sélectionner efficacement dans la masse d'informations, les plus pertinentes, qui seront ensuite interprétées et auxquelles on tachera de donner un sens et évaluer une crédibilité ?

Aujourd'hui, les plateformes de veille tentent de répondre à ces premières problématiques, mais du fait de l'évolution rapide des technologies, des formats et des structures d'information, ces systèmes ne sont pas toujours cohérents et évolutifs. S'il existe de nombreux outils publics et accessibles sur Internet (moteurs de recherche généralistes ou verticaux), une veille qui repose uniquement sur ceux-ci se trouve rapidement limitée. Pour répondre à des besoins d'information précis et garder une réactivité importante face à de nouveaux types d'information, la sélection des techniques les plus pertinentes reste indispensable. De plus, la prise en main de ces outils s'avère coûteuse en temps pour les analystes,

c'est pourquoi l'efficacité d'un travail de veille va dépendre de l'intégration de ces divers outils au sein d'une seule et même plateforme, mais aussi et surtout des performances de chacun des composants de traitement de l'information.

1.2 *Media Mining* & la plateforme WebLab

La coordination d'un ensemble de techniques de traitement de l'information pour les besoins que nous venons d'évoquer est notamment l'objet de recherche de la fouille de documents multimédia ou *Media Mining*. En effet, l'exploitation d'un tel volume d'informations requiert l'automatisation de tout ou partie des traitements d'analyse, d'interprétation et de compréhension des contenus. Il s'agit donc de rechercher, de collecter, d'extraire, de classer, de transformer ou, plus généralement, de traiter l'information issue des documents disponibles et, enfin, de la diffuser de façon sélective en alertant les utilisateurs concernés. Cet ensemble d'analyses est généralement implémenté sous forme d'une même chaine de traitement. Ceci permet de diminuer la quantité d'outils que l'analyste doit maîtriser et utiliser durant sa veille et par là même de faciliter le passage de l'information entre les différents services d'analyse. Ces chaines de traitement apportent une valeur ajoutée dans la recherche d'informations à partir de sources ouvertes mais également dans le traitement de l'ensemble des informations numériques. Dans le cadre du ROSO, elles implémentent des technologies principalement issues des domaines de l'Intelligence Artificielle (IA) et de la gestion des connaissances (KM pour *Knowledge Management*).

Concernant le traitement des données textuelles, par exemple, différentes approches complémentaires peuvent être utilisées. Ainsi, des techniques statistiques permettent d'analyser les contenus d'un grand nombre de documents pour déterminer automatiquement les sujets abordés en fonction des termes les plus discriminants. Par ailleurs, des techniques probabilistes peuvent également être utilisées avec succès pour identifier la langue d'un document. Des techniques d'analyse linguistique à base de grammaires permettent de réaliser d'autres types de traitement en Extraction d'information (EI) tels que la recherche d'expressions régulières, d'amorces de phrases, de noms de personnes, de dates d'événements, etc. Une analyse sémantique permet, quant à elle, de traduire les chaines de caractères ou les données techniques contenues dans les documents multimédia par des concepts de haut niveau. Par ailleurs, des ontologies de domaine peuvent être définies et utilisées pour annoter et rechercher les documents selon un modèle commun bien défini.

La même variété de technologies et approches se retrouve dans le domaine du multimédia. La transcription de la parole, par exemple, permet de réaliser des fonctions similaires aux documents texte sur les documents audio ou vidéo. Une fois transcrit, le document peut être indexé puis retrouvé de façon rapide et efficace. De plus, des outils de traduction peuvent également être intégrés dans une chaîne de traitement. Dans le domaine de l'image, la détection ou la reconnaissance de visages, la recherche par similarité peuvent permettre de retrouver automatiquement une information d'intérêt. La combinaison des techniques de fouille de textes et de fouilles de documents audio ou vidéo ouvre la voie à des traitements de plus en plus puissants. Côté logiciels, il existe un grand nombre de solutions et de briques technologiques mises à disposition par des éditeurs commerciaux ou en *open-source* et par la communauté scientifique. Le choix de la meilleure brique est souvent très difficile car les critères sont nombreux, variés et évolutifs.

Pour composer des offres "sur mesure" bien adaptées au besoin, des plateformes dites d'intégration permettent d'assembler et de faire inter-opérer les outils sélectionnés. Le choix d'une plateforme devient

alors un enjeu essentiel pour produire un système de traitement des informations structurées et non structurées. Seule une plateforme basée sur des standards largement répandus peut permettre d'assurer l'interopérabilité des outils retenus.

WebLab est une de ces plateformes d'intégration, elle est développée et maintenue par la société Cassidian et constitue le socle fonctionnel et technque au sein duquel nos recherches se sont déroulées [Giroux et al., 2008] La plateforme WebLab vise à faciliter l'intégration de composants logiciels plus particulièrement dédiés, au traitement de documents multimédia et d'informations non-structurées (texte, image, audio et vidéo) au sein d'applications dédiées à diverses activités de veille telles que le ROSO mais aussi la veille économique et stratégique. Différents composants spécifiques viennent régulièrement enrichir cette plateforme pour lui offrir des fonctionnalités de collecte de données sur des sources ouvertes (Internet, TV ou radio, presse écrite, etc.) ou dans des entrepôts privés, de traitement automatique des contenus (extraction d'information, analyse sémantique, classification, transcription de la parole, traduction, segmentation, reconnaissance d'écriture, etc.), de capitalisation (stockage, indexation, enregistrement dans des bases de connaissance, etc.) et d'exploitation des connaissances (recherche avancée, visualisation et synthèse graphique, aide à la décision, etc.).

Les objectifs scientifiques et technologiques de la plateforme sont multiples :

- Il s'agit de définir un modèle de référence, basé sur les standards du Web Sémantique (XML, RDF, RDFS, OWL, SPARQL, etc.), permettant à des composants logiciels hétérogènes d'échanger efficacement des données brutes, des méta-données associées ou des informations élaborées de façon automatique.
- Il s'agit également de proposer des interfaces génériques de services afin de normaliser les interactions entre les composants et de simplifier la construction de chaînes de traitement au sein desquelles ils sont mis en œuvre conjointement.
- Il s'agit enfin de proposer et de mettre à disposition un ensemble de briques logicielles réutilisables et composables pour construire rapidement des applications adaptées à un besoin particulier. Ces briques prennent la forme de services qui couvrent un large spectre de fonctionnalités et de composants d'IHM [1] qui permettent de piloter les services et d'en exploiter les résultats côté utilisateur.

Enfin, l'IHM est constituée par assemblage de composants qui s'exécutent au sein d'un portail Web personnalisable par l'utilisateur final. Pour des besoins nouveaux ou spécifiques, les services et composants d'IHM peuvent être créés de toute pièce ou développés en intégrant des composants du commerce et/ou *open-source*. L'architecture de cette plateforme est résumée par la figure 1.

2 Objectifs et axes de recherche

Étant donné le contexte que nous venons de décrire, cette thèse a pour objectif de faciliter et de réduire le travail des analystes dans le cadre du ROSO et de la veille plus généralement. Face à l'augmentation croissante des informations disponibles pour tous librement et légalement, notamment sur le Web, et face à l'hétérogénéité des contenus, il s'agit de proposer un système global de capitalisation des connaissances permettant aux acteurs du ROSO d'exploiter cette masse d'informations.

1. Interface Homme-Machine

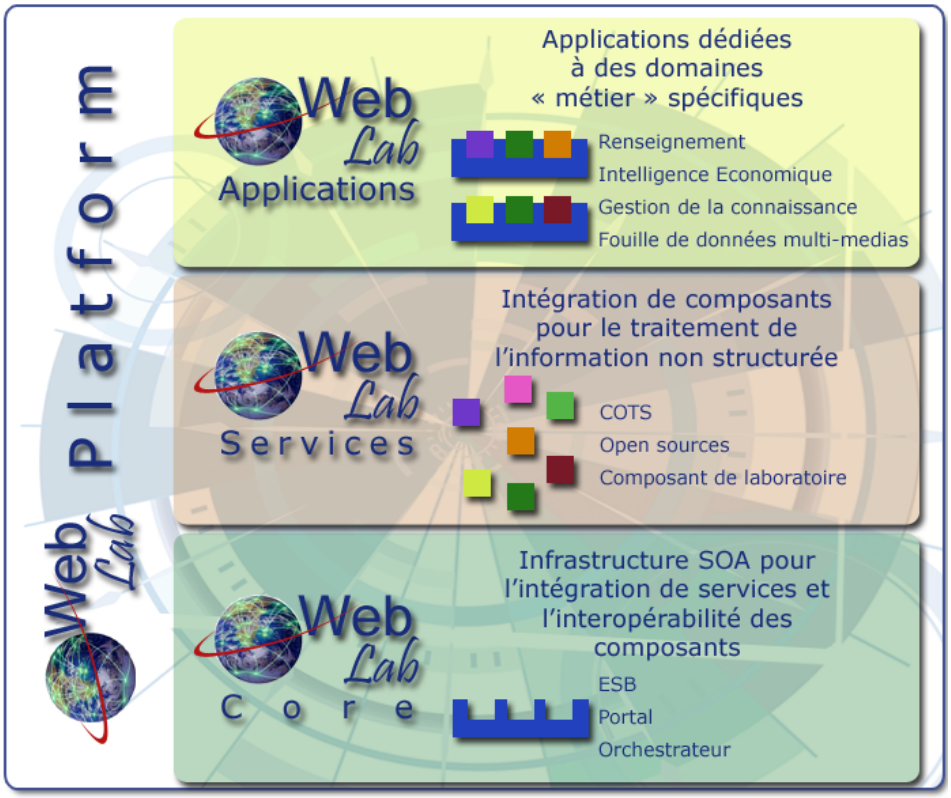


FIGURE 1 – Architecture de la plateforme WebLab

Nos recherches s'articulent autour de trois axes principaux, correspondant à trois des nombreuses problématiques de l'Intelligence Artificielle :

- Modélisation de l'information
- Extraction d'information
- Capitalisation des connaissances

Nos travaux au sein de ce premier axe visent à définir l'étendue et la nature des informations d'intérêt pour le domaine du ROSO, c'est-à-dire mettre en place un modèle de connaissances. Plus concrètement, il s'agira de recenser, définir et formaliser sémantiquement l'ensemble de concepts utilisés par les experts de ce domaine et leurs relations. Ce modèle servira de socle de référence à notre processus global de capitalisation des connaissances : pour exprimer l'ensemble des informations de façon unifiée mais aussi assurer une communication entre les différents services de traitement de l'information développés. Nous explorerons pour cela les travaux existants et notamment les représentations sous forme d'ontologie de domaine qui est, à l'heure actuelle, le mode de représentation le plus utilisé dans ce but.

Le second axe a pour objectif de proposer une approche nouvelle d'extraction d'information à partir de textes en langage naturel. Celle-ci devra permettre de repérer automatiquement l'ensemble des entités et événements d'intérêt pour le ROSO définis au sein du premier axe de recherche. Pour ce faire,

nous nous intéresserons notamment à la combinaison de différentes techniques actuelles (linguistiques, statistiques ou hybrides) afin d'améliorer la qualité des résultats obtenus.

Le dernier axe de recherche vise à définir un processus de transformation des informations extraites en réelles connaissances, c'est-à-dire les normaliser, les structurer, les relier, considérer les problématiques de continuité (redondance/contradiction, temps/espace), etc. Ces traitements doivent aboutir à la création de fiches de connaissances destinées aux analystes, résumant l'ensemble du savoir acquis automatiquement au sujet d'une entité d'intérêt. Celles-ci seront stockées et gérées au sein d'une base de connaissances pour permettre leur mise à jour lors du traitement de nouveaux documents mais également des mécanismes de raisonnement/inférence afin d'en déduire de nouvelles connaissances.

Une place importante sera réservée à l'articulation de ces trois axes de recherche au sein d'un processus global de capitalisation des connaissances que nous souhaitons maintenir le plus générique et flexible possible. La figure 2 présente de façon synthétique la problématique et les objectifs de nos recherches.

3 Contributions de la thèse

Nos travaux de recherche ont donné lieu à plusieurs contributions selon les objectifs que nous venons de définir ainsi qu'à un certain nombre de publications que nous listons ci-dessous.

Nous avons tout d'abord réalisé un état de l'art des différents axes de recherche abordés par le sujet. Suite à cela, nous avons mis en place une ontologie de domaine nommée WOOKIE (Weblab Ontology for Open sources Knowledge and Intelligence Exploitation) dédiée à la représentation des connaissances spécifiques au ROSO et à la veille de façon plus générale. Nous avons notamment défini, et intégré au sein de WOOKIE, un modèle de représentation de l'événement en prenant pour base les conclusions de l'état de l'art. Un événement y est défini comme une entité complexe à quatre dimensions : une dimension conceptuelle (le type de l'événement), une dimension temporelle (la date de l'événement), une dimension spatiale (le lieu de l'événement) et une dimension agentive (les participants de l'événement).

Dans un second temps, nous avons élaboré et évalué un système d'extraction d'événements dit "mixte". En effet, les travaux explorés dans le domaine de l'extraction d'information ayant mis en évidence un certain nombre de limites aux techniques existantes (symboliques et statistiques), nous nous sommes orientés vers une approche combinant deux techniques actuelles. La première méthode proposée consiste en des règles d'extraction élaborées manuellement couplées avec une analyse syntaxique en dépendance. La seconde est basée sur un apprentissage dit "symbolique" de patrons linguistiques par extraction de motifs séquentiels fréquents. Nous avons implémenté ces deux extracteurs en prenant pour base l'ontologie de domaine WOOKIE ainsi que notre représentation des événements et en assurant une possible intégration au sein de la plateforme WebLab. Une première évaluation de ces deux extracteurs a été mise en œuvre et publiée (voir les publications ci-dessous). Chacune des deux méthodes a obtenu des résultats satisfaisants et comparables à l'état de l'art. Cette évaluation a également montré qu'un processus d'agrégation adapté permettra d'exploiter au mieux les points forts de ces deux approches et d'ainsi améliorer significativement la qualité de l'extraction d'événements.

Ce processus d'agrégation constitue notre troisième contribution. Pour cela, nous avons exploré notamment les travaux existants en fusion de données et plus particulièrement en fusion d'informations textuelles. Nous avons choisi d'élaborer un processus d'agrégation sémantique multi-niveaux : une simi-

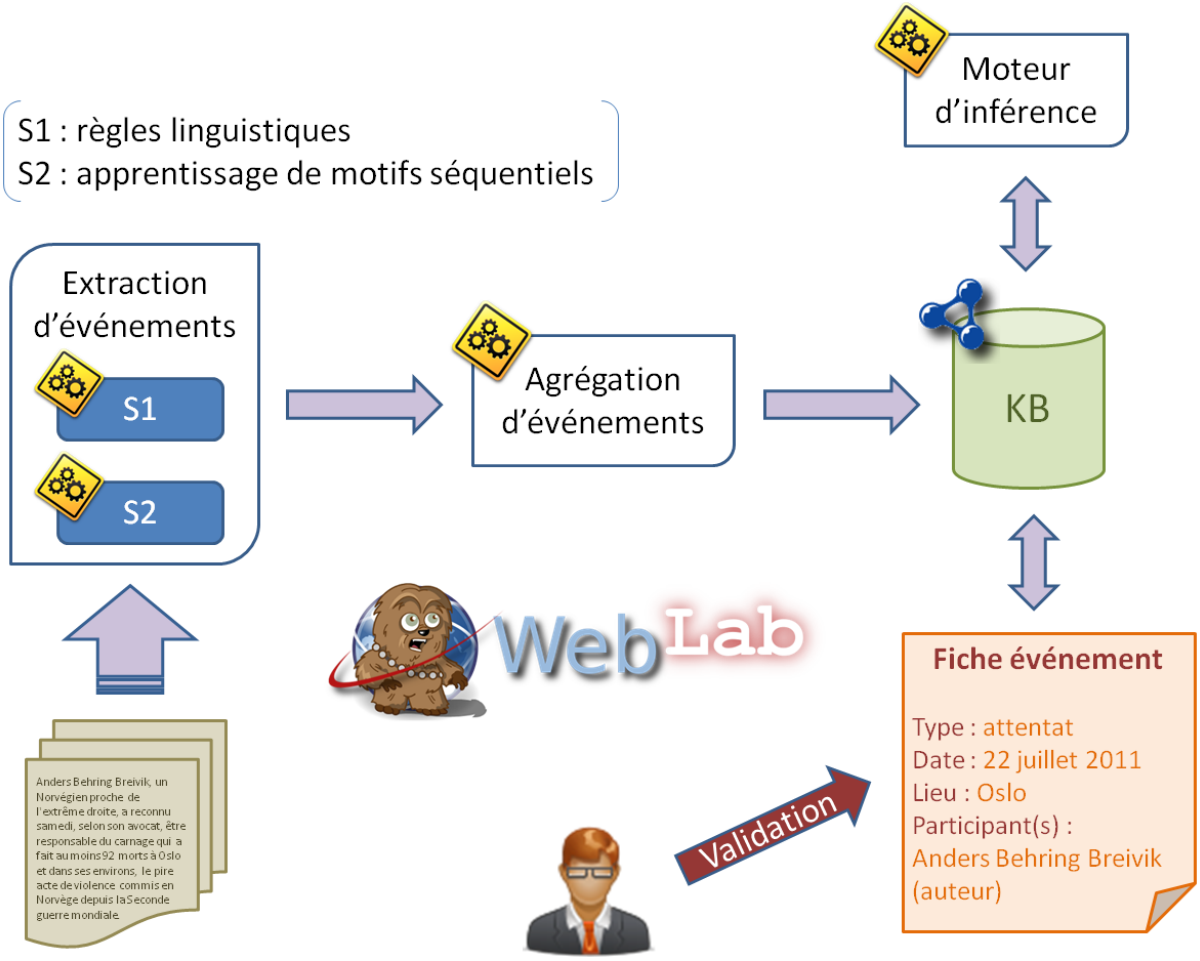


FIGURE 2 – Système de capitalisation des connaissances proposé

larité entre événements est estimée au niveau de chaque dimension puis nous définissons un processus d'agrégation global basé sur ces similarités intermédiaires pour aider l'utilisateur à déterminer si deux événements extraits réfèrent ou non à un même événement dans la réalité. Les similarités sont exprimées selon une échelle qualitative définie en prenant en compte les besoins des utilisateurs finaux. Enfin, nous avons implémenté un prototype d'évaluation permettant l'agrégation des événements suivant le processus.

Nos travaux de recherche ont donné lieu à plusieurs publications dans des conférences nationales et internationales dans les domaines abordés par cette thèse :

- Serrano, L., Grilheres, B., Bouzid, M., and Charnois, T. (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)*, Brest, France
- Serrano, L., Charnois, T., Brunessaux, S., Grilheres, B., and Bouzid, M. (2012b). Combinaison d'approches pour l'extraction automatique d'événements (automatic events extraction by combining multiple approaches) [in french]. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN*, Grenoble, France. ATALA/AFCP
- Serrano, L., Bouzid, M., Charnois, T., and Grilheres, B. (2012a). Vers un système de capitalisation des connaissances : extraction d'événements par combinaison de plusieurs approches. In *Atelier des Sources Ouvertes au Web de Données (SOS-DLWD'2012) en conjonction avec la conférence internationale francophone (EGC 2012)*, Bordeaux, France
- Caron, C., Guillaumont, J., Saval, A., and Serrano, L. (2012). Weblab : une plateforme collaborative dédiée à la capitalisation de connaissances. In *Extraction et gestion des connaissances (EGC'2012)*, Bordeaux, France
- Serrano, L., Bouzid, M., Charnois, T., Brunessaux, S., and Grilheres, B. (2013b). Extraction et agrégation automatique d'événements pour la veille en sources ouvertes : du texte à la connaissance. In *Ingénierie des Connaissances 2013 (IC 2013)*, Lille, France
- Serrano, L., Bouzid, M., Charnois, T., Brunessaux, S., and Grilheres, B. (2013a). Events extraction and aggregation for open source intelligence: from text to knowledge. In *IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2013)*, Washington DC, USA

4 Organisation du mémoire

Ce mémoire est organisé en deux parties reflétant l'ensemble du travail de recherche accompli durant cette thèse :

- la première partie *État de l'art* est divisée en trois chapitres et présente l'étude de l'état de l'art réalisée ;
- la seconde partie *Contributions de la thèse*, composée de quatre chapitres, expose l'ensemble des contributions réalisées.

Le premier chapitre, intitulé *Représentation des connaissances*, propose un tour d'horizon, centré sur notre problématique, du domaine de la représentation et de la modélisation des connaissances. Nous commençons par rappeler succinctement les concepts de base dans ce cadre avant d'aborder la thématique de l'information sémantique avec notamment les travaux autour du Web sémantique et des ontologies. Enfin, nous centrons notre présentation sur l'objet central à cette thèse – les événements – afin de rappeler comment ce concept et ses propriétés ont été définis et modélisés jusqu'à nos jours au sein des différents

axes de recherche mentionnés précédemment.
Le second chapitre *Extraction automatique d'information* réalise un état de l'art autour de la problématique principale de nos travaux, à savoir l'extraction automatique d'information. Dans celui-ci nous recensons notamment les différentes recherches menées récemment autour des trois grands types d'objets de l'EI que sont kes entités nommées, les relations et enfin les événements. Puis, nous nous focalisons sur des aspects plus applicatifs à travers la présentation de quelques plateformes/logiciels pour l'EI et un certain nombre de cas d'application dans ce domaine. Nous clôturons ce chapitre en abordant les problématiques d'évaluation et les performances des méthodes proposées en EI.
Le chapitre *Capitalisation des connaissances* termine cette partie d'état de l'art par une présentation d'ensemble de la problématique nouvelle que constitue la capitalisation des connaissances. Celle-ci aborde tout d'abord les travaux liés à notre sujet de recherche dans des domaines tels que la fusion et la réconciliation de données mais également le mouvement nouveau autour du Web de données. Enfin, nous présentons un ensemble d'approches existantes visant à appliquer les techniques de capitalisation au traitement des événements.

En seconde partie, le quatrième chapitre, intitulé *Modélisation des connaissances du domaine*, détaille la première contribution proposée durant nos travaux de thèse : une modélisation des événements ainsi qu'une ontologie de domaine nommée WOOKIE. Celles-ci ont été élaborées en fonction des conclusions de notre état de l'art et de façon adaptée à notre problématique et notre cadre de recherche, à savoir l'extraction automatique des événements dans le cadre du ROSO.
Le chapitre *Extraction automatique des événements* constitue le cœur de nos recherches et présente notre seconde contribution, c'est-à-dire l'ensemble de nos réalisations autour du second axe de nos recherches. Nous y détaillons, tout d'abord, la méthode que nous avons élaborée pour l'extraction automatique des entités nommées pour les langues anglais et française. Puis, est explicitée et exemplifiée notre contribution centrale, à savoir la conception et la réalisation d'une approche pour l'extraction automatique des événements fondée sur deux méthodes issues de l'état de l'art que nous améliorées et adaptées à notre problématique et à notre cadre applicatif.
Le chapitre suivant *Agrégation sémantique des événements* présente les recherches que nous avons menées dans le cadre du troisième axe "Capitalisation des connaissances". Tout d'abord, nous nous sommes intéressés à la réconciliation de divers méthodes et systèmes pour proposer une méthodologie générique d'agrégation sémantique des événements issus des outils d'extraction. Ce chapitre présente ses fondements autour d'une approche permettant d'estimer la similarité sémantique entre événements et ensuite de les agréger pour faciliter le travail des analystes du ROSO.
Notre mémoire se poursuit avec le dernier chapitre nommé *Expérimentations et résultats* dans lequel sont exposées deux expérimentations réalisées dans le cadre de cette thèse dans le but d'estimer les apports et limites des contributions présentées. La première évaluation concerne notre approche pour l'extraction automatique des événements : pour ce faire nous avons employé un corpus de test issu d'une campagne d'évaluation en EI ainsi que des métriques classiques dans cette discipline. La seconde évaluation est qualitative est montre l'apport de notre méthode d'agrégation sémantique des événements au travers d'exemples réels issus de nos travaux. Pour chacune de ces expérimentations nous exposons également leurs limites ainsi que les perspectives envisagées.

Nous conclurons ce mémoire de recherche en rappelant l'ensemble des contributions réalisées, puis nous exposerons les différentes perspectives ouvertes par nos travaux.

Première partie

État de l'art

Introduction

Cette première partie a pour objectif de réaliser un tour d'horizon de l'existant dans les principaux domaines abordés par cette thèse. Le premier chapitre de cet état de l'art (chapitre 1) sera centré sur les concepts et approches actuels en représentation des connaissances. L'accent sera mis ici sur les technologies du Web sémantique et la modélisation des événements. Le chapitre suivant (chapitre 2) explorera les principales recherches et réalisations dans le domaine de l'extraction d'information et plus particulièrement les travaux concernant l'une de nos principales problématiques, à savoir l'extraction automatique des événements. Pour finir, nous aborderons, au travers du chapitre 3, un ensemble de travaux menés autour de la capitalisation des connaissances, notamment en fusion de données et résolution de coréférence entre événements. Pour conclure chacun de ces trois chapitres, nous dresserons un bilan des forces et faiblesses des approches explorées et nous introduirons chacune de nos contributions en réponse à cet état de l'art.

Chapitre 1

Représentation des connaissances

Sommaire

Dans ce premier chapitre nous réalisons un tour d'horizon, centré sur notre problématique de recherche, des domaines de la représentation et de la modélisation des connaissances. Nous commençons par rappeler succinctement les concepts de base de donnée, information et connaissance, avant d'aborder la thématique de l'information sémantique avec notamment les travaux autour du Web sémantique et des ontologies. Enfin, nous centrons notre présentation sur l'objet central à cette thèse – les événements – afin de déterminer comment ce concept et ses propriétés sont définis et modélisés par les différents travaux actuels.

1.1 Données, informations et connaissances

La distinction entre les termes "donnée", "information" et "connaissance" est couramment abordée dans la littérature liée à la gestion des connaissances [Balmisse, 2002] [Crié, 2003] [Paquet, 2008]. Celle-ci est nécessaire afin, d'une part, de mieux comprendre les différentes problématiques soulevées par le traitement de l'information (au sens large) et, d'autre part, de pointer à quel niveau d'analyse se situent les différents technologies et outils existants.

Une donnée est généralement définie comme un élément brut non traité et disponible hors de tout contexte. Une fois collectées et traitées, par le cerveau humain ou par une machine, ces données deviennent des informations. Une information est le résultat de la contextualisation d'un ensemble de données afin d'en saisir les liens et de leur donner un sens. L'information est statique et périssable, sa valeur diminue dans le temps (car dépendante de son contexte qui est amené à varier). A l'inverse, la connaissance est le résultat d'un processus dynamique visant à assimiler/comprendre les principes sous-jacents à l'ensemble des informations obtenues. Cette compréhension permet de prévoir l'évolution future d'une situation et d'entreprendre des actions en conséquence. Sous réserve de cette interprétation profonde, une plus grande quantité d'information mène à une meilleure connaissance d'un sujet donné.

Prenons par exemple la donnée brute "11/20", sans indication de contexte, cette suite de symboles peut véhiculer diverses informations et ne peut être exploitée en l'état. Toutefois, si cela fait référence à la note obtenue an mathématiques par Pierre, cette donnée contextualisée prend du sens et devient une information. Par ailleurs, si cette première information est associée avec le fait que la moyenne de la classe s'élève à 13/20, chacun peut faire le lien entre ces deux informations et savoir (obtenir la connaissance) que Pierre a des difficultés dans cette matière. Ainsi, cette connaissance acquise, les parents de Pierre pourront, par exemple, prendre la décision de l'inscrire à du soutien scolaire.

La distinction entre ces trois concepts correspond à la dimension hiérarchique de la connaissance proposée par [Charlot and Lancini, 2002]. Les auteurs définissent quatre autres dimensions dont la dimension épistémologique introduisant deux grandes formes de connaissance : celle dite "explicite" qui peut être codifiée et partagée, d'une part, et celle dite "tacite", d'autre part, difficilement exprimable dans un langage commun et donc peu transmissible. Cette définition suggérée par [Polanyi, 1966] est largement partagée par les cogniticiens.

Depuis les débuts du Web, ce passage des données aux connaissances s'est placé au centre des préoccupations de divers secteurs d'activité (industriel, gouvernemental, académique, etc.). En effet, avec l'explosion de la quantité d'information mise en ligne, il est devenu primordial, en particulier pour les organisations, de pouvoir exploiter cette masse afin d'en obtenir des connaissances. Nous sommes passés ainsi de l'époque des bases de données à celle des bases de connaissances.

1.2 L'information sémantique

Le développement du Web et des NTIC [2] a mis en avant un nouvel enjeu : le partage des connaissances. D'un Web de documents (Web 1.0) voué à la publication/visualisation statique des informations grâce au langage HTML, nous avons évolué vers un Web plus collaboratif et dynamique. Ce Web 2.0, introduit aux débuts des années 2000, a constitué une première évolution dans ce sens en remettant l'homme au centre de la toile grâce aux blogs, réseaux sociaux, wikis et autres moyens lui permettant de créer, publier et partager ses propres connaissances. Dès les début du Web 2.0, Tim Berners-Lee évoquait déjà la prochaine "version" du World Wide Web : le Web sémantique [Berners-Lee et al., 2001]. Les sections suivantes présentent ce Web 3.0 et les différentes technologies associées.

1.2.1 Le Web sémantique

Le Web sémantique désigne un projet d'évolution de notre Web actuel (Web 2.0) initié par le W3C [3]. Cette initiative est aussi connue sous les noms de Web de données, *Linked Data*, *Linked Open Data* ou encore *Linking Open Data*. L'objectif principal de ce mouvement est de faire en sorte que la quantité de données exposée sur le Web (qui ne cesse de croître) soit disponible dans un format standard, accessible et manipulable de manière unifiée par toutes les applications du Web.

Le Web sémantique de Tim Berners-Lee est voué à donner du sens à l'ensemble des données mises en ligne afin de faciliter la communication entre les hommes et les machines et permettre ainsi à ces dernières de réaliser des tâches qui incombaient jusqu'alors aux utilisateurs. Par ailleurs, le nom Web de données met en avant la nécessité de créer des liens entre données pour passer d'un Web organisé en silos de données déconnectés (ayant leur propres formats, protocoles, applications, etc.) à un espace unifié sous la forme d'une collection de silos inter-connectés.

La figure 1.1 présente l'état actuel du LOD [4] sous la forme d'un graphe où les noeuds correspondent à des bases de connaissances respectant les principes du Web sémantique et les arcs représentent les liens existants. Parmi les plus renommés des silos de données, la base sémantique DBPedia fournit une grande partie du contenu de Wikipedia et incorpore des liens vers d'autres bases de connaissances telles que Geonames, par exemple. Ces relations (sous la forme de triplets RDF [5], voir la section 1.2.3) permettent aux applications Web d'exploiter la connaissance supplémentaire (et potentiellement plus précise) issue d'autres bases de connaissances et de fournir ainsi de meilleurs services à leurs utilisateurs.

2. Nouvelles Technologies de l'Information et de la Communication
3. World Wide Web Consortium, `http://www.w3.org/`
4. Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. `http://lod-cloud.net/`
5. Resource Description Framework, `http://www.w3.org/RDF/`

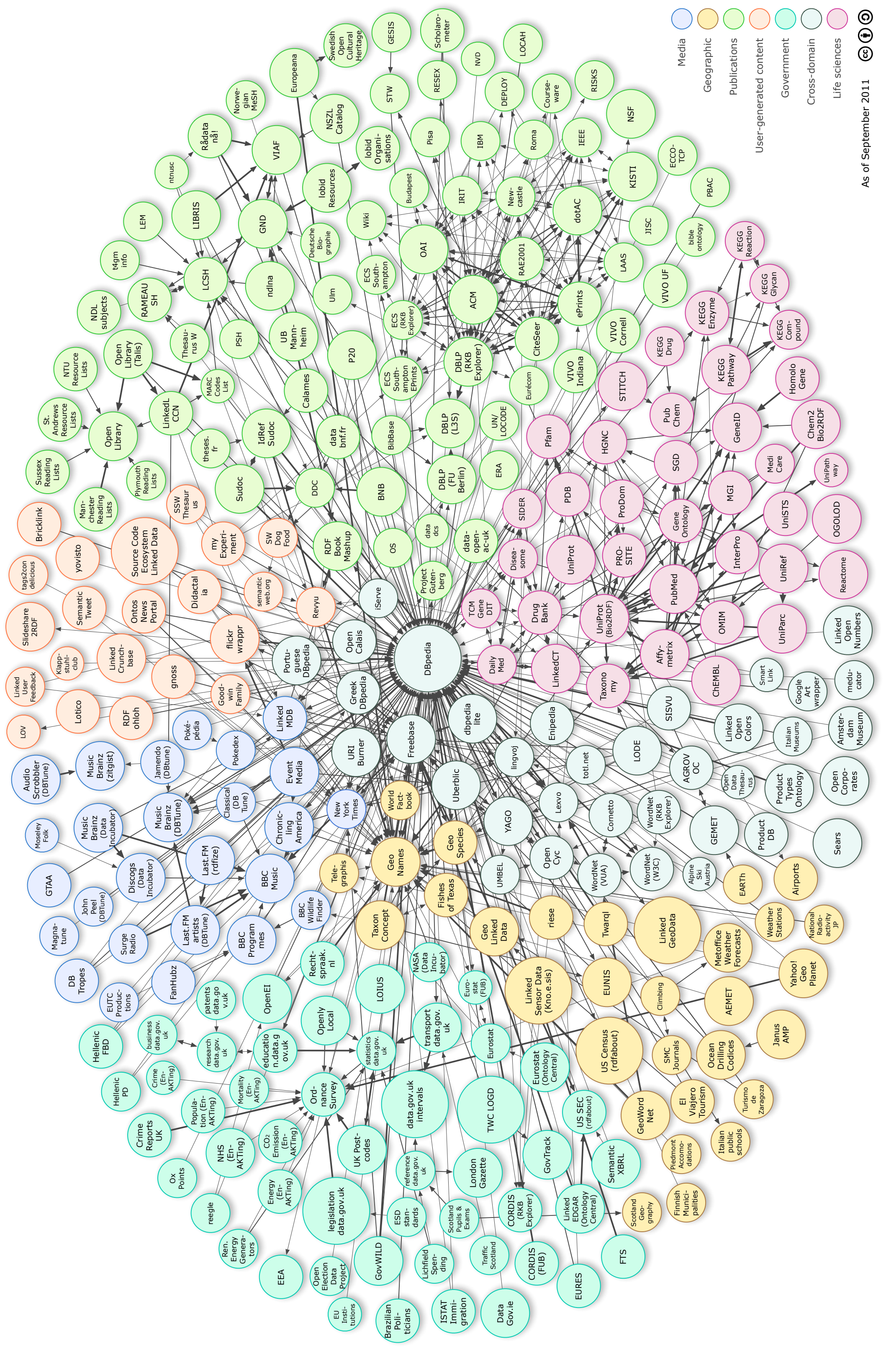


FIGURE 1.1 – Linking Open Data

Pour mettre en œuvre les principes du Web sémantique, le W3C recommande un ensemble de technologies du Web sémantique (RDF, OWL, SKOS, SPARQL, etc.) fournissant un environnement dans lequel les applications du Web peuvent partager une modélisation commune, interroger "sémantiquement" le LOD (Linked Open Data), inférer de nouvelles connaissances à partir de l'existant, etc.

1.2.2 Les ontologies

Le concept d'ontologie s'avère aujourd'hui indissociable du Web sémantique. Toutefois, les ontologies ne sont pas nouvelles et sont les héritières des travaux en Ontologie (la science de l'Être) menés par des philosophes de le Grèce antique tels qu'Aristote ou Platon. L'ontologie telle qu'on l'entend en informatique est maintenant assez éloignée de ces études menées au carrefour entre la métaphysique et la philosophie.

Plusieurs définitions ont été proposées dont celles de [Gruber, 1993] et [Neches et al., 1991]. Le premier donne une définition très abstraite des ontologies mais largement admise dans la littérature : "Une ontologie est une spécification explicite d'une conceptualisation" [6]. Le second propose celle-ci : "Une ontologie définit les termes et les relations de base du vocabulaire d'un domaine ainsi queMaître de conférences les règles qui permettent de combiner les termes et les relations afin de pouvoir étendre le vocabulaire" [7]. La définition de [Gruber, 1993] renvoie à deux caractéristiques principales des ontologies à savoir, d'une part, l'élaboration d'un modèle abstrait de l'existant (conceptualisation) et, d'autre part, sa formalisation en vue d'une exploitation par des machines (spécification explicite).

Les ontologies à l'ère du Web sémantique sont aux bases de connaissances ce que les modèles de données sont aux bases de données [Charlet et al., 2004]. Elles définissent l'ensemble des objets du domaine de connaissances ciblé ainsi que les attributs et relations caractérisant ces objets. Les objets sont représentés sous forme de concepts hiérarchisés constituant la taxonomie de classes de l'ontologie. Cette relation de subsomption se retrouve dans toutes les ontologies, quelque soit le domaine concerné : les classes de plus haut niveau dans la hiérarchie correspondent à des concepts généraux et les classes inférieures représentent des concepts plus spécifiques. S'ajoutent à cette taxonomie des relations entre concepts de nature diverse ainsi que des attributs de concepts. Les premières ont pour domaine ("domain" en anglais) et co-domaine ("range" en anglais) des classes de l'ontologie tandis que les seconds ont pour domaine une classe et pour co-domaine une valeur d'un certain type (chaine de caractères, nombre, date, booléen, etc.). Ce type de modélisation implique un phénomène d'héritage : chaque classe de l'ontologie hérite des propriétés (relations et attributs) de sa classe supérieure (dite "classe mère"). Pour résumer, nous pouvons dire qu'une ontologie définit sémantiquement et de façon non-ambigüe un ensemble de concepts et propriétés issus d'un consensus.

Un exemple commun dans la communauté d'ingénierie des connaissances est l'ontologie des *pizza.owl* servant couramment de base aux tutoriels dédiés aux ontologies. Cet exemple-type très complet contient un ensemble de classes, propriétés, attributs et instances du domaine ainsi que des fonctionnalités plus avancées de contraintes, cardinalités et axiomes. L'annexe G présente une vue d'ensemble de cette ontologie grâce au logiciel Protégé (présenté en section 1.2.5).

6. traduction de l'anglais "An ontology is an explicit specification of a conceptualization"
7. traduction de l'anglais "An ontology defines the basic terms and relations comprising the vocabulary of a topic area as well as the rules for combining terms and relations to define extensions to the vocabulary"

Par ailleurs, on distingue différents types d'ontologie selon la portée de la modélisation [Guarino, 1998] : les ontologies dites "générales" ou "de haut niveau", les ontologies de domaine, de tâche ou d'application. Les premières visent à décrire les objets du monde communs à plusieurs domaines tels que le temps et l'espace alors que les trois autres types modélisent des concepts spécifiques. La création d'ontologies de haut niveau est parfois vue comme une utopie et la majorité des travaux de la littérature se fondent sur des ontologies de domaine.

Plusieurs aspects peuvent rendre la construction d'une ontologie délicate et coûteuse en temps. En effet, ce travail de modélisation comporte une part de subjectivité car plusieurs visions d'un même domaine sont possibles et le temps pour arriver à un consensus peut s'en trouver allongé. Ce phénomène s'amplifie avec la complexité du domaine à modéliser mais également avec la taille de la communauté concernée. Afin de faciliter la tâche d'élaboration d'une ontologie des travaux tels que [Noy and Mcguinness, 2001], [Mizoguchi, 2003a] ou encore [Mizoguchi, 2003b] suggèrent un ensemble de bonnes pratiques.

1.2.3 Les langages de représentation

Pour permettre aux ordinateurs d'exploiter cette modélisation, des langages informatiques de spécification d'ontologie ont été créés dans le cadre du Web Sémantique.

Les premiers langages ont été développés par la DARPA au début des années 2000, il s'agit de DAML [8], OIL [9] [Fensel et al., 2001] ou DAML+OIL [Horrocks, 2002]. Ceux-ci sont les ancêtres des langages RDFS et OWL [10] devenus les recommandations actuelles du W3C. La majorité d'entre eux est fondée sur des formalismes inspirés du modèle des assertions en logique de premier ordre. RDF est le formalisme recommandé par le W3C mais d'autres existent tels que Common Logic [11] [Bachmair and Ganzinger, 2001], DOGMA [12] [Jarrar and Meersman, 2009], KIF [13] [Genesereth, 1991], F-Logic [14] [Kifer et al., 1995], etc.

RDF est un formalisme pour la représentation de faits sur le Web fondé sur la notion de triplet. Tel les assertions en logique des prédicats, un triplet RDF est composé de trois éléments : un sujet, un prédicat et un objet. Le sujet et le prédicat sont des ressources et, dans le cas du Web sémantique, il s'agit de tout objet pouvant être identifié (une page Web, une personne, une propriété, etc.). Une ressource Web est représentée par un URI [15] qui peut être, par commodité, raccourci grâce à un espace de nom (*namespace*). L'objet du triplet, quant à lui, peut être soit une ressource (le triplet exprime une relation entre objets) soit une valeur (le triplet exprime un attribut d'un objet). L'ensemble des valeurs possibles en RDF est emprunté du format XML. Bien que le W3C recommande l'utilisation du formalisme RDF/XML, d'autres types de sérialisation existent telles que les formats N3 (Notation3), N-triples, Turtle, JSON, etc.

8. DARPA Agent Markup Language
9. Ontology Inference Layer
10. Ontology Web Language, `http://www.w3.org/TR/owl-features/`
11. `http://iso-commonlogic.org/`
12. Developing Ontology-Grounded Methods and Applications, `http://www.starlab.vub.ac.be/research/dogma.htm`
13. Knowledge Interchange Format, `http://www.ksl.stanford.edu/knowledge-sharing/kif/`
14. Frame Logic
15. Uniform Resource Identifier

Le langage RDFS est une première extension de RDF visant à structurer les ressources pour la spécification des ontologies. Les propriétés les plus utilisées du formalisme RDFS sont les suivantes : *rdfs :Class*, *rdfs :subClassOf*, *rdfs :domain*, *rdfs :range* et *rdfs :label*. Les deux premières servent à définir la taxonomie de l'ontologie, les deux suivantes permettant de formaliser la notion de sujet et d'objet et enfin, la propriété *label* sert à nommer les éléments de l'ontologie (classes et propriétés).

Enfin, le langage largement privilégié aujourd'hui pour la modélisation des ontologies est OWL. Développé depuis 2002 par un groupe de travail du W3C, celui-ci constitue une seconde extension plus expressive du standard RDF/XML. OWL permet notamment grâce à des constructeurs d'exprimer des contraintes supplémentaires sur les classes et propriétés définies : disjonction, union et intersection de classes, propriétés symétriques, inverses, etc. Une majeure partie de ces constructeurs est directement issue des logiques de description permettant des mécanismes d'inférence sur les connaissances de l'ontologie. Précisons également que le langage OWL se décline en plusieurs sous-langages selon leurs niveaux d'expressivité : OWL-Lite, OWL-DL et OWL-Full. Nous présentons en annexe H un extrait de l'ontologie *pizza.owl* exprimée au format OWL.

1.2.4 Inférence et bases de connaissances

Nous pouvons définir l'inférence comme le fait de déduire de nouvelles connaissances par une analyse de l'existant. Dans le cas des ontologies, il s'agit concrètement de découvrir de nouveaux triplets à partir des triplets connus en exploitant la structure de l'ontologie et les contraintes logiques spécifiées. Autrement dit, l'inférence permet de faire apparaître automatiquement des faits implicites que l'œil humain ne peut détecter car ils sont masqués par la complexité de la modélisation. Ce genre de raisonnement peut être effectué au niveau de l'ontologie en elle-même (*terminological box* ou *TBox*) ou au niveau des instances de l'ontologie (*assertional box* ou *ABox*). Il faut noter que plus le langage de représentation utilisé est expressif, plus l'inférence sera complexe voire impossible à mettre en œuvre. Dans le cas du langage OWL, la complexité de l'inférence est croissante lorsque l'on passe du sous-langage OWL-Lite à OWL-DL et enfin à OWL-Full. Ces mécanismes d'inférence sont implémentés dans des raisonneurs tels que Pellet [16], Fact++ [17] ou encore Hermitt [18]. Lorsque la complexité de la modélisation est trop élevée pour ces raisonneurs, il est possible de définir des règles d'inférence manuellement grâce à des langages tels que SWRL [19] ou *Jena rules* [20].

Ces mécanismes de déduction constituent une réelle plus-value des ontologies car ils permettent, d'une part, de vérifier la qualité des bases de connaissance construites et, d'autre part, d'y ajouter de nouvelles connaissances de façon entièrement automatique. Les systèmes de gestion de base de données classiques (MySQL, Oracle Database, IBM DB2, etc.) ont vu se créer leurs équivalents sémantiques, nommés *triplestores*, permettant le stockage et le requêtage de triplets RDF. Les grands fournisseurs de SGBD propriétaires adaptent aujourd'hui leurs solutions au Web sémantique et des *triplestores open-source* sont également disponibles comme Sesame [21], Virtuoso [22] ou encore Fuseki [23]. La récupération

16. http://clarkparsia.com/pellet/
17. http://owl.man.ac.uk/factplusplus/
18. http://www.hermit-reasoner.com/
19. Semantic Web Rule Language, http://www.w3.org/Submission/SWRL/
20. http://jena.apache.org/documentation/inference/
21. openRDF, http://www.openrdf.org/
22. OpenLink, http://virtuoso.openlinksw.com/
23. Apache Jena, http://jena.apache.org/

de ces données sémantiques est réalisée majoritairement grâce au langage SPARQL [24], l'équivalent sémantique du langage SQL [25] dont l'usage est recommandé par le W3C pour l'interrogation de bases de connaissances sémantiques. La requête ci-dessous, par exemple, permet de récupérer l'ensemble des entités de type *foaf :Person* dont le nom de famille commence par la lettre *S*.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT ?f ?l WHERE
{
  ?p rdf:type foaf:Person .
  ?p foaf:firstName ?f .
  ?p foaf:lastName ?l .
  FILTER(regex(?l, "^S"))
}
```

1.2.5 Les éditeurs d'ontologies

Bien que l'intérêt pour les ontologies sur le Web soit relativement récent, de nombreux outils ont été développés dans le but de modéliser et de manipuler des ontologies. Le principal atout de ces logiciels est la possibilité de gérer une ontologie dans l'un des formats cités précédemment sans avoir à modifier manuellement le code sous-jacent. Nous présentons ici quelques logiciels distribués librement et gratuitement.

Tout d'abord, SWOOP [26] est un éditeur simplifié d'ontologies *open-source*, développé par l'université du Maryland. Implémenté en Java, cet outil supporte les formats RDF (différentes sérialisations) et OWL. Parallèlement à ses fonctions d'édition, Swoop permet d'effectuer des raisonnements et propose un service de recherche des ontologies existantes.

D'autre part, OntoWiki [27] est une application Web conçue comme un wiki permettant de gérer une ontologie de manière simple et collaborative. Cet outil est développé par le groupe de recherche AKSW [28] de l'université de Leipzig, également connu pour leur projet DBPedia. Cet outil supporte plusieurs formats tels que RDF/XML, Notation3, Turtle ou encore Talis(JSON). L'accent est également mis sur l'aspect *Linked Data* et l'intégration de ressources externes. Des extensions sont disponibles pour par exemple attacher des pages de wiki à des ressources de l'ontologie, visualiser des informations statistiques grâce à CubeViz ou encore intégrer des fonds de carte géographiques.

L'outil TopBraid Composer [29] est fourni par la société anglo-saxonne TopQuadrant et s'insrcit dans la suite d'outils professionnels TopBraid Suite. Plusieurs versions de TopBraid Composer sont disponibles dont une édition gratuite Free Edition (FE). Cette application est implémentée en Java grâce à la plateforme Eclipse et exploite les fonctionnalités de la librairie Apache Jena. Elle supporte l'import et

24. SPARQL Protocol and RDF Query Language, http://www.w3.org/TR/sparql11-overview/
25. Structured Query Language
26. http://code.google.com/p/swoop/
27. http://ontowiki.net/Projects/OntoWiki
28. Agile Knowledge engineering and Semantic Web
29. http://www.topquadrant.com/products/TB_Composer.html

l'export de fichiers en langage RDF(S), OWL et OWL2 et son système de built-ins ouvre de nombreuses autres fonctionnalités telles que l'utilisation de divers moteurs d'inférence, le requêtage en SPARQL, le développement de règles d'inférence en SWRL, les vérification d'intégrité et gestion des exceptions.

L'application Apollo [30] est développée par le Knowledge Media Institute (KMI). Cet outil implémentée en Java fournit les fonctions basiques de création d'ontologie et présente une bonne flexibilité grâce à son système de *plug-ins*. Toutefois, Apollo ne permet pas de mécanismes d'inférence et est fondé sur son propre métalangage ce qui ne facilite pas l'interopérabilité avec d'autres outils.

NeOn Toolkit [31] est un autre environnement *open-source* très complet d'ingénierie des ontologies issu du projet européen NeOn. Cet outil est particulièrement adapté à la gestion d'ontologies multi-modulaires ou multilingues, à la fusion d'ontologies et à leur intégration dans des applications sémantiques plus larges. L'accent y est mis sur la contextualisation et la mise en réseau de modélisations hétérogènes et sur les aspects collaboratifs de la gestion d'ontologies. Fondé sur l'environnement de développement Eclipse, NeOn Toolkit propose l'intégration de divers plugins pour la modélisation visuelle, l'apprentissage et l'alignement d'ontologie, la définition de règles d'inférence, etc. Son architecture ouverte et modulaire est compatible avec les architectures orientées services (SOA) [Haase et al., 2008].

Terminons par l'éditeur d'ontologies le plus renommé et le plus utilisé dans la communauté d'ingénierie des connaissances : l'environnement Protégé [32]. Créé par les chercheurs de l'université de Stanford, Protégé est développé en Java, gratuit et *open-source*. Il s'agit d'une plateforme d'aide à la création, la visualisation et la manipulation d'ontologies dans divers formats de représentation (RDF, RDFS, OWL, etc.). Ce logiciel peut également être utilisé en combinaison avec divers moteurs d'inférence (tels que RacerPro ou Fact) afin d'effectuer des raisonnements et d'obtenir de nouvelles assertions. De plus, de par la flexibilité de son architecture, Protégé est facilement configurable et extensible par les *plugins* développés au sein d'autres projets. La figure 1.2 présente une vue de l'ontologie-exemple *pizza.owl* dans l'environnement Protégé. Enfin, les créateurs de cet outil mettent l'accent sur l'aspect collaboratif dans la modélisation d'ontologies en proposant Collaborative Protégé et WebProtégé. Le premier est une extension intégrée à Protégé permettant à plusieurs utilisateurs d'éditer la même ontologie et de commenter les modifications effectuées par chacun. Un système de vote rend également possible la concertation sur tel ou tel changement. WebProtégé est une application Web légère et *open-source* reprenant les principes de Collaborative Protégé dans le contexte du Web. Elle permet une édition d'ontologies collaborative et à distance.

Pour une description plus détaillée et une comparaison plus poussée de ces éditeurs, nous renvoyons le lecteur aux présentations faites par [Alatrish, 2012] et [Charlet et al., 2004].

1.3 Modélisation des événements

Le concept d'événement étant au centre de nos travaux, les sections suivantes visent à définir plus précisément cette notion complexe. Nous nous intéressons aux différentes visions de l'événement dans la littérature et plus particulièrement dans les domaines de l'extraction d'information et du Web sémantique. Nous résumons les grands courants de représentation des événements dans le but de les extraire

30. http://apollo.open.ac.uk/
31. http://neon-toolkit.org
32. http://protege.stanford.edu/

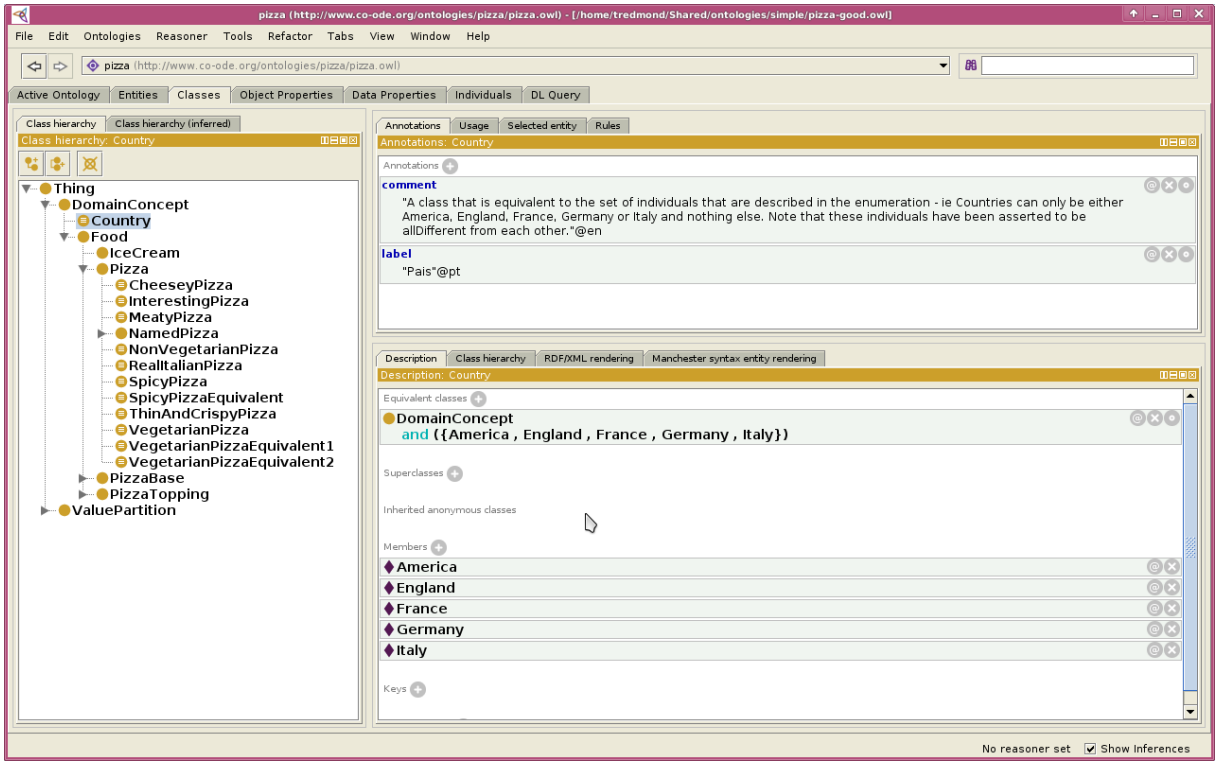


FIGURE 1.2 – L'environnement Protégé

automatiquement. Puis, nous réalisons un état des lieux des ontologies existantes centrées sur la modélisation des événements. Enfin, un focus est proposé sur les questions de représentation du temps et de l'espace qui sont des aspects importants dans le traitement des événements.

1.3.1 Qu'est-ce qu'un événement ?

Considéré comme une entité aux propriétés bien spécifiques, l'événement a initialement été étudié par des philosophes [Davidson, 1967] puis par des linguistes [Van De Velde, 2006] [Desclés, 1990] (voir [Casati and Varzi, 1997] pour une bibliographie exhaustive).

[Higginbotham et al., 2000] compare quelques définitions de l'événement et met notamment l'accent sur deux courants traditionnels opposés : les événements comme concepts universaux (généraux et répétables) d'une part, et particuliers (spécifiques et uniques), d'autre part. Dans le premier courant, on peut citer [Montague, 1969] pour lequel les événements sont des propriétés attribuées à des instants (ou des intervalles) dans le temps, ou encore [Chisholm, 1970] qui les définit comme des situations ("states of affairs") pouvant décrire le même intervalle de temps de façon différente. Par ailleurs, [Quine, 1960] et [Kim, 1973] appartiennent au second courant : les événements sont définis comme des objets particuliers ("individuals"). Le premier considère que les événements ont le même statut que les objets physiques et portent le contenu (parfois hétérogène) d'une portion de l'espace-temps. Il n'y a donc qu'un seul et même événement possible par région spatio-temporelle mais ce même événement peut donner lieu à différentes descriptions linguistiques. [Kim, 1973] se place à l'opposé de cette vision en permettant à un nombre

indéfini d'événements d'occuper un seul et même moment. L'événement est ici un objet concret (ou une collection d'objets) exemplifiant une propriété (ou un ensemble de propriétés) à un moment donné. Enfin, une position intermédiaire est celle de [Davidson, 1969] qui considère les événements selon leur place dans un réseau de causalité : des événements sont identiques s'ils partagent les mêmes causes et les mêmes effets. Cependant, ce même auteur reviendra plus tard vers la thèse de Quine.

Suite à ces réflexions sur la nature des événements (qui est encore débattue à l'heure actuelle), d'autres recherches ont vu le jour en linguistique, analyse du discours et philosophie du langage. En effet, les événements étant avant tout des objets sociaux, leur étude doit se faire entre étroite corrélation avec la manière dont ils sont relatés et exprimés par l'homme. Parmi ces travaux nous pouvons citer [Krieg-Planque, 2009]. Celle-ci donne une définition simple de l'événement mais qui nous paraît très juste : "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 est reconnue comme partie intégrante de ce concept par la quasi totalité des travaux. 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.". C'est ici qu'est fait le lien avec la sociologie et l'histoire : tout événement prend place dans un milieu social qui détermine l'obtention de son statut "remarquable".

Enfin, [Neveu and Quéré, 1996] s'attachent à décrire plus précisément l'apparition des événements dits "modernes", façonnés et relayés par les médias actuels. 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 appellerons ces termes des "noms d'événement". Ces "noms d'événement" transposent en langage naturel la "propriété sémantique" des événements mentionnée par [Saval, 2011]. Cette description de l'événement est également au centre d'un phénomène plus large, que [Ricœur, 1983] nomme "mise en intrigue" : celui-ci vise à organiser, selon le cadre mentionné plus haut, un ensemble d'éléments circonstants ou participants de l'événement.

1.3.1.1 Les événements en extraction d'information

On distingue actuellement deux grandes visions de l'événement dans la communauté de l'extraction d'information [Ahn, 2006]. D'une part, l'approche TimeML provient de recherches menées en 2002 dans le cadre du programme AQUAINT [33] fondé par l'ARPA [34]. D'autre part, le modèle ACE a été défini dans la tâche Event Detection and Recognition (VDR) des campagnes d'évaluation du même nom à partir de 2005.

L'approche TimeML [Pustejovsky et al., 2003] définit les événements de la façon suivante : "situations that happen or occur [...] predicates describing states or circumstances in which something obtains or holds true". Ceux-ci peuvent être ponctuels ou duratifs et sont organisés en sept types : *Occurrence*, *State*, *Reporting*, *I-Action*, *I-State*, *Aspectual*, *Perception*. L'événement est considéré conjointement à trois autres types d'entité : *TIMEX*, *SIGNAL* et *LINK*. Les objets *TIMEX* correspondent aux entités temporelles simples telles que les expressions de dates et heures, durée, fréquence, etc. (annotées par des étiquettes *TIMEX3* [36]). Les entités *SIGNAL* correspondent à des mots fonctionnels exprimant des rela-

33. Advanced Question Answering for Intelligence, http://www-nlpir.nist.gov/projects/aquaint/
34. ancien nom de la DARPA, Defense Advanced Research Projects Agency, [35]
36. http://timeml.org/site/timebank/documentation-1.2.html

tions temporelles. Il peut s'agir de prépositions de temps (pendant, après, etc.), de connecteurs temporels (quand, lorsque, etc.), de mots subordonnants (si, etc.), d'indicateurs de polarité (négation) ou encore de quantifieurs (dix fois, souvent, etc.). Enfin, les annotations *LINK* font le lien entre les différentes entités temporelles (*EVENT*, *TIMEX* et *SIGNAL*). Ces liens sont de trois types : *TLINK* (liens temporels entre événements ou entre un événement et une autre entité de type *TIMEX*), *SLINK* (liens de subordination entre deux événements ou entre un événement et une entité *SIGNAL*), *ALINK* (liens aspectuels entre un événement aspectuel et son argument).

Dans le modèle ACE [NIST, 2005] un événement est vu comme une structure complexe impliquant plusieurs arguments pouvant également être complexes. L'aspect temporel correspond ici à un type d'argument mais n'est pas au centre de la modélisation. Ce modèle définit un ensemble de types et sous-types d'événement (8 types comme *Life*, *Movement*, *Business*, etc. et 33 sous-types comme *Marry*, *Declare-Bankruptcy*, *Convict*, *Attack*, etc.) et associe à chaque événement un ensemble d'arguments autorisés pouvant être des entités ou des valeurs et auxquels est associé un rôle (parmi 35 rôles dont Time, Place, Agent, Instrument, etc. [LDC, 2005]). Par ailleurs, les événements tels que définis dans la campagne ACE possèdent des attributs tels que le temps, la modalité, la polarité ou encore la généricité. Enfin, comme pour les autres types d'entité, on distingue la notion de mention d'événement, d'une part, qui correspond à une portion de texte constituée d'une ancre d'événement et de ses arguments et, d'autre part, l'événement en lui-même, c'est-à-dire un ensemble de mentions d'événement qui réfère au même objet du monde réel.

Ces deux modèles sont différents sur plusieurs aspects, la divergence principale étant que la première approche vise à annoter tous les évènements d'un texte, alors que la seconde a pour cibles uniquement les événements d'intérêt pour une application donnée. Le modèle TimeML, considérant un événement comme tout terme temporellement ancré, est généralement choisi dans des projets où cet aspect est central (pour la construction de chronologies par exemple). L'approche ACE est la plus utilisée car elle s'applique aux besoins de divers domaines mais implique toutefois un processus d'annotation plus complexe à mettre en place (la représentation de l'événement étant elle-même plus complexe).

1.3.1.2 Les ontologies orientées "événement"

Plusieurs ontologies disponibles sur le Web proposent une modélisation spécifique au concept d'événement. Nous présentons ci-après les caractéristiques principales de cinq d'entre elles :

- The Event Ontology (EO) [37]
- Linking Open Descriptions of Events (LODE) [38]
- Simple Event Model (SEM) [39]
- DOLCE-UltraLite (DUL) [40]
- CIDOC Conceptual Reference Model (CIDOC CRM) [41]

The Event Ontology a été développée par les chercheurs Yves Raimond et Samer Abdallah du *Centre for Digital Music* à Londres [Raimond et al., 2007] et sa dernière version date d'octobre 2007. Cette

37. http://motools.sourceforge.net/event/event.html
38. http://linkedevents.org/ontology
39. http://semanticweb.cs.vu.nl/2009/11/sem/
40. http://ontologydesignpatterns.org/ont/dul/DUL.owl
41. cidoc-crm.org/

ontologie est centrée sur la notion d'événement telle que [Allen and Ferguson, 1994] la définissent : "[...] events are primarily linguistic or cognitive in nature. That is, the world does not really contain events. Rather, events are the way by which agents classify certain useful and relevant patterns of change.". Cette ontologie des événements à été développée dans le cadre du projet *Music Ontology* et est donc particulièrement adaptée à la représentation des événements dans ce domaine (concerts, représentations, etc.). Comme le montre la figure 1.3, cette modélisation se veut générique et s'appuie sur des ontologies largement utilisées telles que FOAF, ainsi que sur des standards de représentation spatio-temporelle recommandés par le W3C [42]. Toutefois, la classe *Event* n'est pas sous-typée et les propriétés *factor* et *product* paraissent vaguement définies (pas de co-domaine).

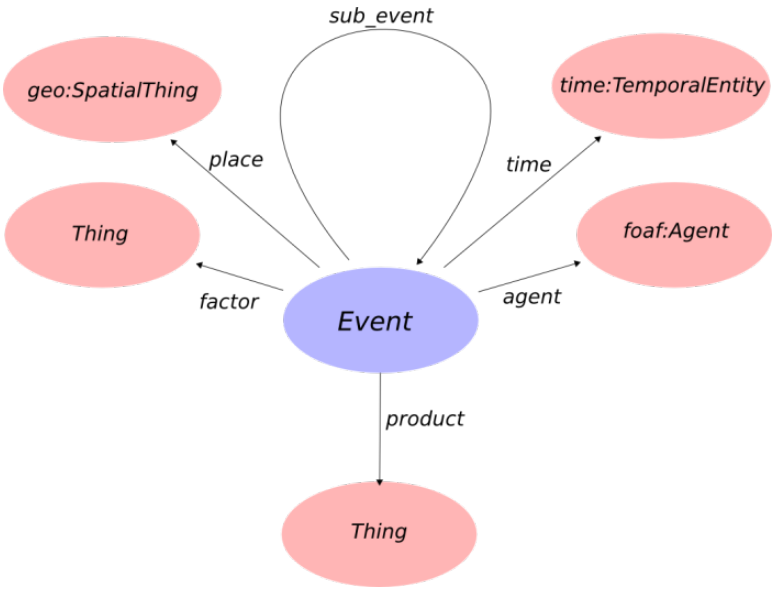


FIGURE 1.3 – L'ontologie Event : modélisation des événements

L'ontologie LODE [Troncy et al., 2010] a été créée dans le cadre du projet européen *EventMedia* [43] [Fialho et al., 2010]. Ce projet vise à concevoir un environnement Web permettant aux internautes d'explorer, de sélectionner et de partager des événements. L'ontologie est au centre de cette initiative car elle constitue le socle commun pour représenter et stocker des événements provenant de sources hétérogènes. Les concepteurs de cette ontologie la présente comme un modèle minimal dont l'objectif premier est, dans la mouvance du LOD (*Linked Open Data*), de créer des liens entre les ontologies existantes afin de représenter les aspects faisant consensus dans la communauté de représentation des événements. Pour ce faire, un certain nombre d'alignements entre ontologies est implémenté dans LODE tels que des équivalences de classes et propriétés avec les ontologies DUL, EO, CIDOC CRM, etc. Cette interopérabilité permet notamment d'obtenir des connaissances supplémentaires lors des mécanismes d'inférence. Cette ontologie n'est donc pas réellement une ontologie des événements mais plutôt un outil d'alignement des modélisations existantes. Enfin, les concepteurs ont exclu pour le moment tout travail sur la sous-catégorisation des événements et la définition de relations entre événements (inclusion, causalité, etc.). Les figures 1.4 et 1.5 schématisent respectivement les relations entre concepts et les relations entre propriétés dans LODE.

42. http://www.w3.org/2006/time# et http://www.w3.org/2003/01/geo/wgs84_pos#
43. http://eventmedia.cwi.nl/

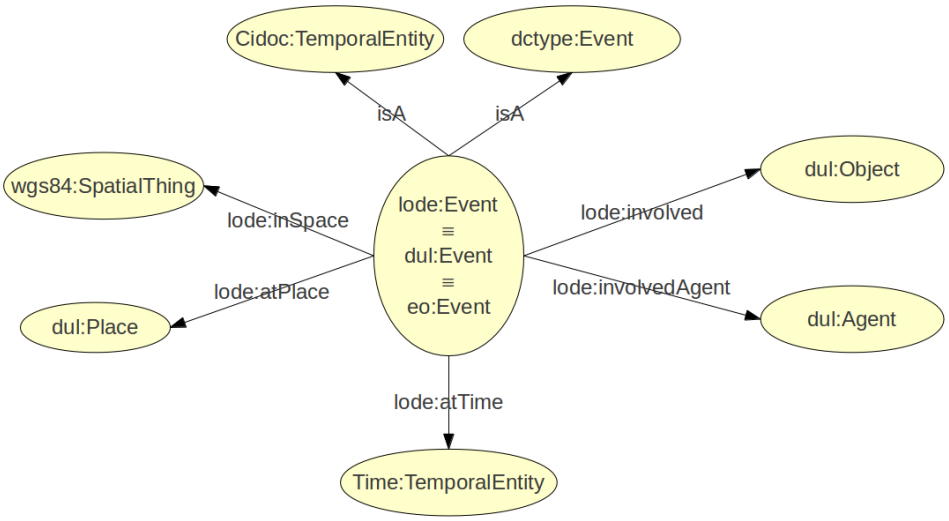


FIGURE 1.4 – LODE : modélisation des événements

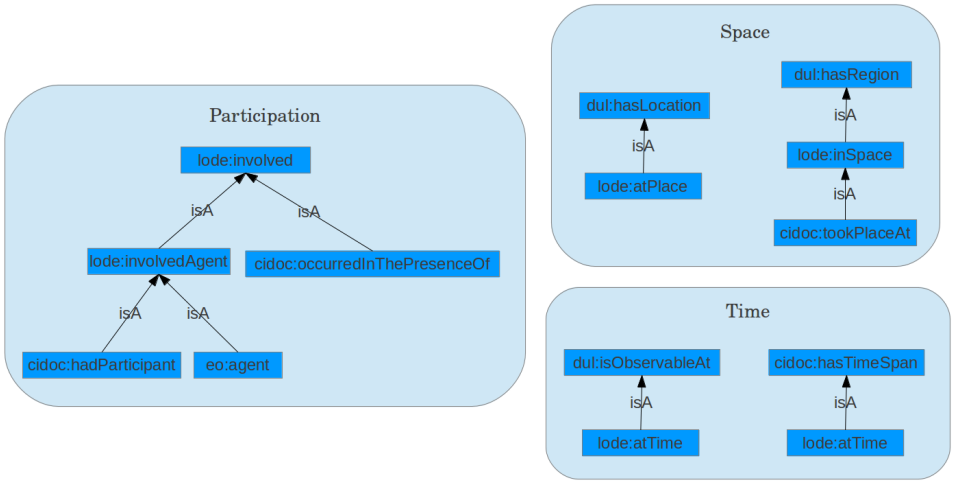


FIGURE 1.5 – LODE : alignements entre propriétés

L'ontologie SEM a été créée par le groupe *Web & Media* de l'université d'Amsterdam [van Hage et al., 2011]. Elle décrit les événements dans le but de répondre à la question "who did what with what to whom, where and when ?" mais aussi dans une perspective d'interopérabilité entre différents domaines. Une particularité de SEM est l'accent mis sur la représentation des rôles associés aux acteurs impliqués dans un événement. SEM permet de modéliser la nature du rôle, des informations temporelles sur sa validité mais également la source l'ayant attribué à l'acteur. Dans le même objectif que LODE, SEM fournit de nombreux liens (sous la forme de propriétés SKOS [44]) avec une dizaine d'ontologies existantes : ontologies dédiées aux événements (EO, LODE, etc.), standards W3C (Time, WGS84), ontologies de haut niveau reconnues (SUMO, OpenCyc, etc.), etc. Toutefois, les événements dans cette ontologie ne sont pas sous-typés et, mise à part la propriété *hasSubEvent*, aucune autre relation entre événements n'est mo-

44. Simple Knowledge Organization System, http://www.w3.org/2004/02/skos/

délisée. La figure 1.6, fournie par les développeurs de l'ontologie, schématise les concepts et relations principaux dans SEM.

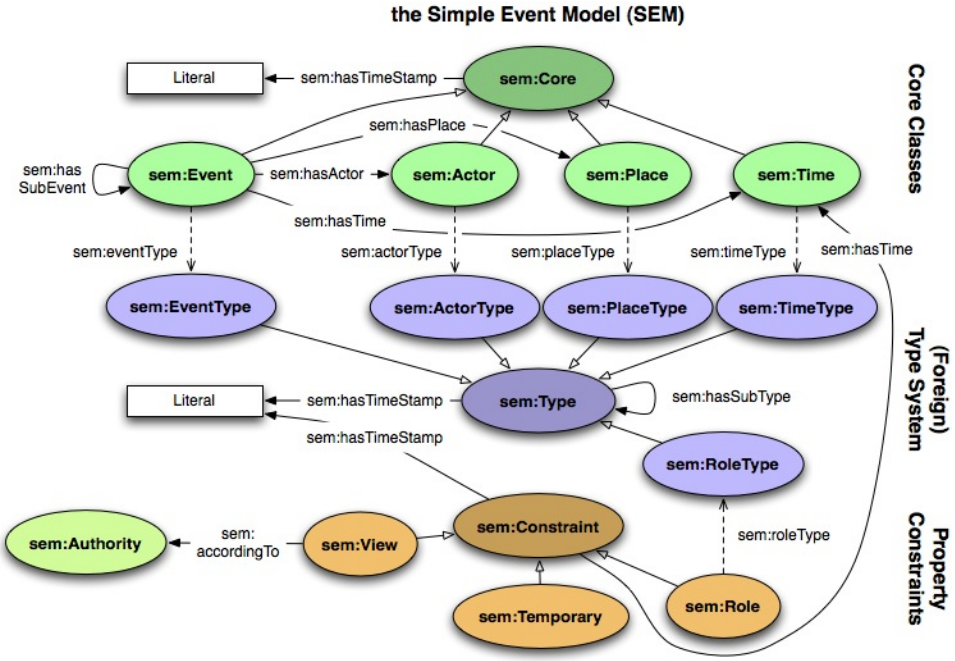


FIGURE 1.6 – SEM : modélisation des événements

DUL (DOLCE+DnS-UltraLite) est une ontologie développée par le laboratoire italien *STLab* [45] à la suite de DOLCE [46], une ontologie de haut niveau créée dans le cadre du projet *WonderWeb* [47]. Un événement y est défini comme suit : "Any physical, social, or mental process, event, or state.". Plus concrètement, la classe *Event* est spécifiée en deux sous-classes *Action* et *Process*, la première se différenciant de la seconde par la propriété *executesTask* indiquant l'objectif de l'action. Comme le montre la figure 1.7, cette ontologie de haut niveau ne comporte pas de liens vers d'autres modélisations et notamment la représentation spatio-temporelle des événements y reste peu définie.

Pour finir, le modèle CIDOC CRM est un standard international ISO pour le partage d'information dans le domaine du patrimoine culturel. Celui-ci est développé depuis 1994 et distribué au format OWL par les chercheurs de l'Université d'Erlangen-Nuremberg en Allemagne [48]. L'événement est une sous-classe de *TemporalEntity* et y est défini comme un changement d'état : "changes of states in cultural, social or physical systems, regardless of scale, brought about by a series or group of coherent physical, cultural, technological or legal phenomena.". La figure 1.8 présente une vue globale de l'organisation des différents concepts de cette ontologie avec un focus sur la classe *Event*. Les différentes propriétés de l'événement sont schématisées dans la figure 1.9. Les remarques faites sur DUL (absence de liens vers d'autres ontologies et représentation spatio-temporelle peu définie) s'appliquent également à cette ontologie. Par ailleurs, la distinction entre les concepts *Event*, *Period* et *Condition State* n'est pas clairement

45. Semantic Technology laboratory
46. Descriptive Ontology for Linguistic and Cognitive Engineering
47. http://wonderweb.semanticweb.org/
48. http://erlangen-crm.org

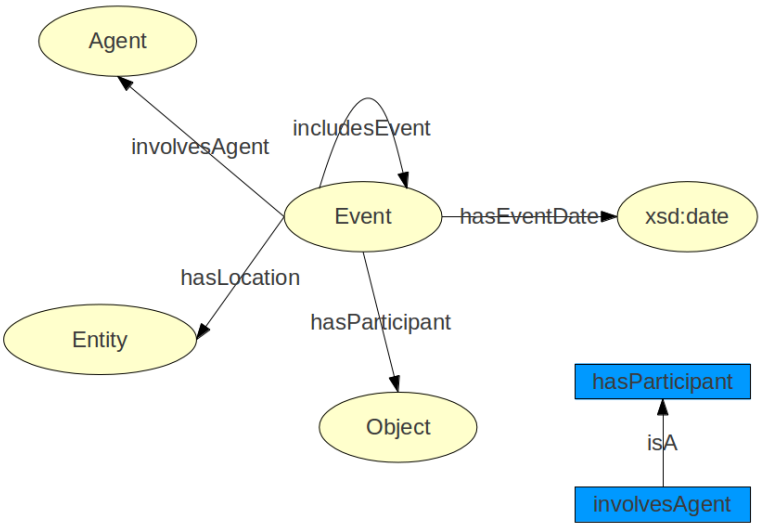


FIGURE 1.7 – DUL : modélisation des événements

définie. Enfin, bien que CIDOC CRM se veuille générique et définisse de nombreuses sous-classes de l'événement, cette taxonomie ne paraît pas s'appliquer à tous les domaines.

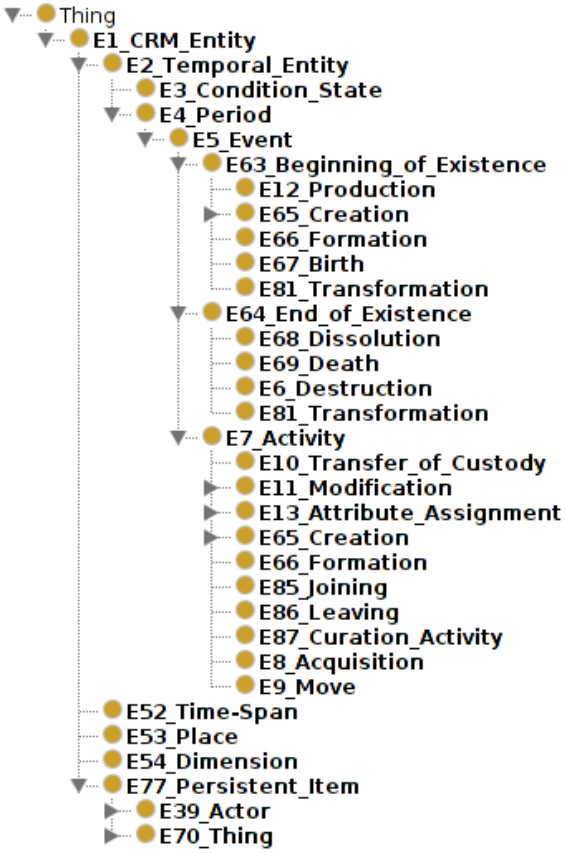


FIGURE 1.8 – CIDOC CRM : taxonomie des classes

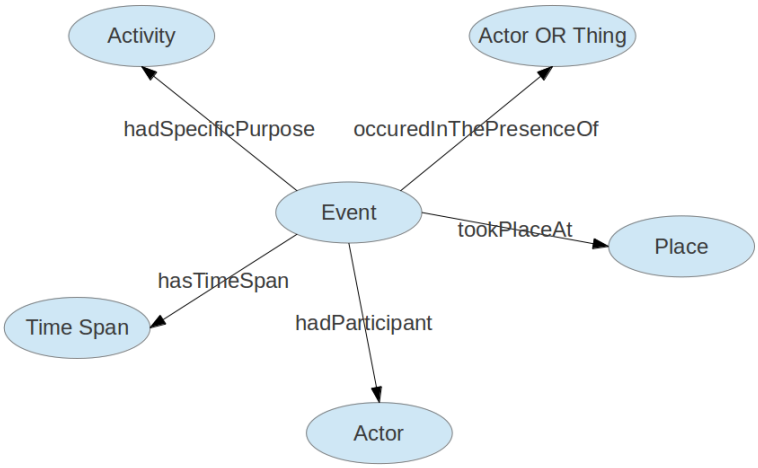


FIGURE 1.9 – CIDOC CRM : modélisation des événements

1.3.2 Modélisation du temps et de l'espace

Comme l'ont montré les sections précédentes, la définition des événements est indissociable des notions de temps et d'espace. Bien que ces concepts soient intuitivement compréhensibles par tous, leur modélisation en vue de traitements automatisés n'est pas chose aisée. En effet, le temps et l'espace peuvent revêtir différentes dimensions sociologiques et physiques donnant lieu à diverses représentations. Nous proposons ci-après un résumé des grands courants théoriques de représentation temporelle et spatiale ainsi que quelques exemples d'ontologies associées.

1.3.2.1 Représentation du temps

On distingue classiquement deux types de représentation temporelle : d'une part, une vue sous forme de points/instants et, d'autre part, un découpage du continuum temporel en intervalles/périodes. Les travaux les plus connus sont ceux de [McDermott, 1982] pour la représentation en points et ceux de [Allen, 1983] pour les intervalles. Le premier considère l'espace temps comme une succession de points et définit les relations suivantes entre deux points t_i et t_j :

$$(t_i < t_j) \vee (t_i > t_j) \vee (t_i = t_j)$$

Allen définit l'unité temporelle de base comme étant un intervalle de temps et propose une algèbre composée de 13 relations (voir la figure 1.10). Ces deux modèles temporels sont les plus utilisés mais d'autre types ont été proposés pour, par exemple, hybrider ces deux approches. De nombreuses recherches sont menées également dans diverses disciplines (philosophie, informatique théorique, bases de données, etc.) sur d'autres problématiques liées au temps telles que les logiques et les contraintes temporelles, le raisonnement sur le temps, etc. [Allen, 1991] [Hayes, 1995]

Parmi ces travaux, [Ladkin, 1987] propose un système de représentation du temps faisant le lien entre les nombreux modèles théoriques développés et les besoins applicatifs de traitement du temps. Il s'agit

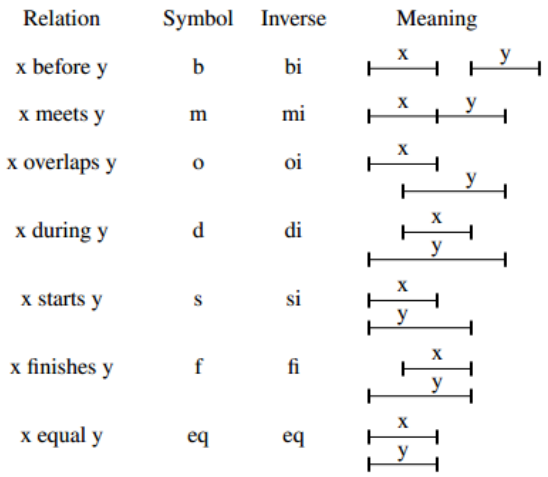


FIGURE 1.10 – Algèbre temporel d'Allen

du *Time Unit System* (TUS), 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*.

Depuis les débuts du Web sémantique, les chercheurs se sont intéressés à l'application des modèles théoriques existants pour la construction d'ontologies temporelles. L'ontologie *OWL Time* [49], issue d'un groupe de travail du W3C, est la plus connue et la plus utilisée actuellement. Celle-ci définit un concept de base *TemporalEntity*, spécifié en instants et intervalles. On y trouve également les relations temporelles issues de l'algèbre d'Allen ainsi qu'un découpage calendaire du temps (année, mois, jour, heure, minute et seconde).

1.3.2.2 Représentation de l'espace

Du côté de la représentation spatiale, les premières approches proviennent des mathématiques : d'une part, Euclide définit plusieurs catégories d'objets géométriques de base (points, lignes, surfaces, etc.) ainsi que leurs propriétés et relations ; d'autre part, Descartes considère le point comme élément fondamental et propose un modèle numérique de représentation géographique en associant à chaque point un ensemble de valeurs pour chacune de ses dimensions (coordonnées cartésiennes). Ces deux modélisations constituent la vision classique de l'espace en termes de points.

49. http://www.w3.org/TR/owl-time/

Plus récemment, se développe une vision alternative où l'unité de base de modélisation spatiale est la région. Introduite par [Whitehead, 1920], celle-ci se fonde sur une perception plus commune de l'espace et est également désignée sous le nom de "géométrie du monde sensible". Cette représentation dite aussi qualitative (en contraste avec les premières approches quantitatives) permet de décrire toute entité spatiale selon les trois concepts suivants : intérieur, frontière et extérieur. Ce principe est à l'origine de l'analyse topologique visant à décrire les relations spatiales selon cette notion de limite. Cette proposition de Whitehead a donné lieu à l'élaboration de diverses théories de l'espace fondées sur les régions dont [Tarski, 1956]. Ce mode de représentation a été celui adopté par la communauté de l'Intelligence Artificielle parce qu'étant plus proche de la vision humaine et donc plus applicatif. Le modèle RCC-8 de [Randell et al., 1992] est le formalisme de premier ordre le plus utilisé pour modéliser et raisonner sur des entités spatiales. Il s'agit d'une transposition de l'algèbre d'Allen au problème de la représentation spatiale. Ce modèle comprend huit types de relation entre deux régions spatiales schématisés par la figure 1.11.

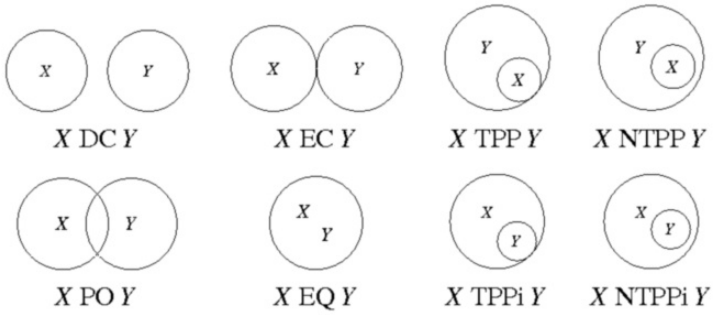


FIGURE 1.11 – Les relations topologiques RCC-8

Du côté du Web sémantique, [Minard, 2008] propose un état de l'art de quelques ontologies d'objets géographiques développées notamment par des laboratoires spécialisés tels que l'INSEE, l'IGN ou AGROVOC. S'ajoutent à ces ontologies de nombreux thésaurus et bases de connaissances géographiques telles que la plus renommée GeoNames [50]. Il nous faut mentionner également le groupe de travail W3C Geospatial Incubator Group (GeoXG) [51] et le consortium international OGC [52] qui collaborent depuis quelques années pour faciliter l'interopérabilité entre les systèmes d'information géographique (SIG) dans le cadre du Web sémantique. Plusieurs initiatives en découlent dont le vocabulaire WGS84 Geo Positioning [53], le langage de balisage GML [54] ou encore le langage de requête GeoSPARQL [55].

1.3.3 Spécifications dédiées au ROSO

Nous proposons ici un tour d'horizon des modélisations réalisées et exploitées dans le domaine militaire et de la sécurité au sens large.

50. http://www.geonames.org
51. http://www.w3.org/2005/Incubator/geo/XGR-geo-ont-20071023/
52. Open Geospatial Consortium, http://www.opengeospatial.org/
53. http://www.w3.org/2003/01/geo/wgs84_pos#
54. Geography Markup Language, http://www.opengeospatial.org/standards/gml
55. http://geosparql.org/

Une première catégorie de ressources existantes sont les standards OTAN [56] dits STANAGs [57]. Ceux-ci sont des accords de normalisation ratifiés par les pays de l'Alliance pour faciliter les interactions entre leurs armées. Ces documents définissent des procédures, termes et conditions formant un référentiel commun (technique, opérationnel ou administratif) dédié à être mis en œuvre au sein des différents services militaires de l'OTAN. Certains sont spécifiques aux systèmes d'information et de communication (SIC) tels que les *NATO Military Intelligence Data Exchange Standard* [NATO, 2001] et *Joint Consultation Command Control Information Exchange Data Model* [NATO, 2007]. Le premier, également nommé STANAG 2433 ou AINTP-3(A), a pour objectif l'échange de l'information et de l'intelligence au sein de l'Alliance et la proposition d'un ensemble de standards d'implémentation. Les structures de données qui y sont définies constituent un exemple pour les pays membres qui, pour la plupart, s'en sont inspirés dans la conception de leurs bases de données. Le second accord, dit STANAG 5525 ou JC3IEDM, sert d'interface commune pour l'échange d'informations sur le champ de bataille. Il vise une interopérabilité internationale entre les systèmes d'information de type C2 (*Command and Control*) pour faciliter le commandement des opérations interarmées. Ces standards militaires ne sont pas des ontologies mais définissent un ensemble de concepts (et relations entre ces concepts) spécifiques à certaines procédures. Ceux-ci sont organisés de façon hiérarchique et décrits le plus précisément possible en langage naturel. Ces modèles sont généralement très complets et très bien documentés car ils sont destinés à des opérateurs humains spécialistes du sujet traité.

Par ailleurs, nous pouvons citer l'agence de recherche américaine IARPA [58] et son programme OSI [59] dont l'objectif est le développement de méthodes pour l'analyse des données accessibles publiquement. Cette organisation finance des projets d'innovation technologique pour la détection automatique et l'anticipation d'événements sociétaux tels que les situations de crise, catastrophes naturelles, etc. Dans le cadre de ce programme a été développée la typologie IDEA [60] incluant environ 250 types d'événements sociaux, économiques et politiques ainsi que des entités simples avec lesquelles ils sont en relation [Bond et al., 2003].

Du côté des ontologies, on recense notamment les modélisations *swint-terrorism* [Mannes and Golbeck, 2005], reprenant les concepts principaux nécessaires au domaine du terrorisme, et *AKTiveSA* [Smart et al., 2007], dédiée à la description des contextes opérationnels militaires autres que la guerre. [Baumgartner and Retschitzegger, 2006] réalise également un état de l'art des ontologies de haut niveau dédiées à la tenue de situation (*situation awareness*). [Inyaem et al., 2010b] s'attache à la définition d'une ontologie floue pour l'extraction automatique d'événements terroristes. Toutefois, comme dans beaucoup de travaux de ce domaine, l'ontologie développée n'est pas distribuée publiquement pour des raisons stratégiques et/ou de confidentialité. Le site web `http://militaryontology.com` recense un certain nombre d'ontologies pour le domaine militaire.

Enfin, [Bowman et al., 2001] et [Boury-Brisset, 2003] proposent un ensemble de conseils méthodologiques pour la construction d'ontologies dédiées aux applications militaires et stratégiques.

56. Organisation du Traité de l'Atlantique Nord
57. STANdardization AGreement, `http://www.nato.int/cps/en/natolive/stanag.htm`
58. Intelligence Advanced Research Projects Activity, `http://www.iarpa.gov/`
59. Open Source Indicators, `http://www.iarpa.gov/Programs/ia/OSI/osi.html`
60. Integrated Data for Events Analysis, `http://vranet.com/IDEA.aspx`

1.4 Conclusions

A travers ce premier état de l'art, nous avons pu nous familiariser avec les différentes problématiques de la représentation des connaissances au sens large. Nous avons, dans un premier temps, rappelé une distinction importante antre les notions de donnée, information et connaissance qui constitue l'une de bases théoriques de ce domaine. Nous avons, par la suite, présenté, dans leur ensemble, les principes et technologies du Web sémantique qui sont partie intégrante d'une majorité de travaux actuels en fouille de documents et dont l'adéquation à cette problématique n'est plus à prouver. Enfin, nous avons réalisé un focus sur la place des événements en représentation des connaissances et présenté les différentes théories et modèles de la littérature. Dans un souci d'interopérabilité avec les systèmes existants, nous avons privilégié l'utilisation de modèles communs au sein de nos travaux tout en les adaptant aux besoins spécifiques de notre application si nécessaire. Le modèle ACE, très utilisé par la communauté en EI, nous parait bien adapté à la modélisation des événements dans le cadre de nos recherches. En effet, comme nous l'avons montré dans ce chapitre, celui-ci est compatible avec les besoins des analystes du ROSO ainsi qu'avec le modèle de données de la plateforme WebLab. Nous avons par la suite réalisé un tour d'horizon des ontologies centrées sur les événements déjà développées et réutilisables : les plus communes présentent toutes des similarités de structure et un effort est réalisé dans la communauté pour lier les différentes modélisations entre elles via des alignements ontologiques. Parmi celles-ci, les ontologies DUL et LODE présentent le plus de correspondances avec nos travaux. L'ensemble de ces observations seront prises en compte lors de l'élaboration de notre modèle de connaissances présentée au chapitre 4.

Chapitre 2

Extraction automatique d'information

Sommaire

Depuis les débuts du Traitement Automatique du Langage dans les années 60-70, la compréhension automatique de textes est l'objet de nombreuses recherches. L'objectif principal est de permettre à un ordinateur de comprendre le sens global d'un document comme savent le faire les êtres humains. Les échecs récurrents des systèmes alors développés mettent rapidement en cause une vision trop générique de la compréhension automatique. En effet, de tels outils s'avèrent alors inutilisables dans un contexte opérationnel en raison du coût élevé des adaptations nécessaires (bases de connaissances et ressources lexicales spécifiques). Conscients d'être trop ambitieux au regard des possibilités technologiques, les chercheurs s'orientent alors vers des techniques plus réalistes d'extraction d'information. S'il n'est pas directement possible de comprendre automatiquement un texte, le repérage et l'extraction des principaux éléments de sens apparaît comme un objectif plus raisonnable. Cette réorientation théorique est reprise de façon détaillée par [Poibeau, 2003].

Ce chapitre présente tout d'abord les objectifs principaux de l'extraction d'information (section 2.1), puis une synthèse des méthodes les plus couramment mises en œuvre dans ce domaine (section 2.2). Pour cela, nous distinguons les travaux existants selon la nature des informations extraites : entités nommées, relations entre entités et événements. Quelques travaux en résolution de coréférence sont également présentés. La section 2.3 propose un tour d'horizon des outils et plateformes existants pour le développement de systèmes d'extraction d'information. Nous parcourons ensuite (section 2.4) quelques-unes des applications possibles dans ce domaine. Pour conclure ce chapitre, la section 2.5 aborde le problème de l'évaluation en extraction d'information à travers une revue des campagnes d'évaluation, puis une présentation des performances (atouts et faiblesses) des systèmes existants et enfin, un récapitulatif des limites restantes à l'heure actuelle.

2.1 Définition et objectifs

Face à l'augmentation vertigineuse des documents textuels mis à disposition de tous, l'extraction automatique d'information voit un intérêt grandissant depuis une vingtaine d'années. En effet, noyés sous cette masse d'information non-structurée, nous rêvons d'un système automatique capable, dans nos tâches quotidiennes (professionnelles ou personnelles), de repérer et d'extraire de façon rapide et efficace les informations dont nous avons besoin. En réponse à cela, les systèmes développés visent à analyser un texte de manière automatique afin d'en extraire un ensemble d'informations jugées pertinentes [Hobbs and Riloff, 2010]. Il s'agit généralement de construire une représentation structurée (bases de données, fiches, tableaux) à partir d'un ou plusieurs documents à l'origine non-structurés. 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 [Poibeau, 2003].

L'extraction d'information (EI) a souvent été confondue avec un autre domaine de l'intelligence artificielle (IA) qu'est la recherche d'information (RI). Bien que cet amalgame s'explique car ces deux champs de recherche partagent un objectif premier — présenter à l'utilisateur des informations qui répondent à son besoin — ceux-ci diffèrent sur plusieurs autres points. Tout d'abord, les outils de RI renvoient généralement une liste de documents à l'utilisateur alors qu'en extraction d'information les résultats sont des éléments d'information extraits de ces documents. Par ailleurs, la recherche d'information s'attache à répondre à une requête exprimée par un utilisateur grâce à un système de mots-clés (ou d'autres mécanismes plus sophistiqués de recherche sémantique) [Vlahovic, 2011]. Dans le domaine de l'EI, la réponse des outils est guidée par une définition *a priori* des éléments d'information à repérer

dans un ensemble de textes. Malgré ces distinctions et comme c'est le cas avec d'autres disciplines du TALN [61], l'EI est utile à la RI et inversement. Les systèmes d'extraction d'information bénéficient souvent de la capacité de filtrage des outils de RI en focalisant leur analyse sur un ensemble déterminé de documents. A l'inverse, la recherche d'information peut exploiter les informations extraites par les outils d'EI en tant que champs de recherche additionnels et ainsi améliorer le filtrage et l'ordonnancement des documents pour une requête donnée.

Les tâches les plus communes en extraction d'information sont l'extraction d'entités nommées [Nadeau and Sekine, 2007], le repérage de relations entre ces entités [Rosario and Hearst, 2005] et la détection d'événements [Naughton et al., 2006]. Celles-ci se distinguent par la nature et la complexité des informations que l'on cherche à repérer et à extraire automatiquement : entités nommées, relations ou événements. Nous détaillons ci-après les objets d'étude et les objectifs spécifiques à chaque tâche.

La reconnaissance d'entités nommées (REN) vise à reconnaître et catégoriser automatiquement un ensemble d'éléments d'information qui correspondent généralement à des noms propres (noms de personnes, organisations, lieux) mais aussi aux dates, unités monétaires, pourcentages, unités de mesure, etc. Ces objets sont communément appelés "entités nommées" et s'avèrent indispensables pour saisir le sens d'un texte. Le terme "entité nommée" (EN) n'est apparu en EI que très récemment lors de la 6ème édition des campagnes d'évaluation MUC [62] (voir la section 2.5) et sa définition est encore aujourd'hui l'objet de nombreuses discussions. Ce terme renvoie à la théorie de la "référence directe" évoquée dès la fin du 19ème siècle par des philosophes du langage tels que John Stuart Mill. A l'heure actuelle, une majorité de travaux se retrouvent dans la définition proposée par [Kripke, 1980] sous le nom de "désignateurs rigides" : "entités fortement référentielles désignant directement un objet du monde". Il nous faut également souligner que ces entités nommées dites "entités simples" constituent généralement un premier niveau de la chaîne d'extraction et permettent la détection de structures plus complexes que sont les relations et les événements. A cette première tâche d'extraction est couramment associé le problème de résolution de coréférence entre EN. La résolution de coréférence vise à regrouper plusieurs extractions ayant des formes de surface différentes mais référant à la même entité du monde : par exemple, "Big Apple" et "New York", ou encore "JFK" et "John Fitzgerald Kennedy". Ce problème a surtout été exploré conjointement à la détection d'entités nommées mais cette problématique est applicable à tout type d'extraction. Dans les campagnes MUC, la résolution de coréférence fait partie d'un ensemble de tâches nommé SemEval (Semantic Evaluation) ayant pour objectif une compréhension plus profonde des textes [Grishman and Sundheim, 1996]. De plus, la différentiation entre les termes "mention" et "entité" introduite lors des campagnes ACE [63] (voir la section 2.5) met en avant ce besoin de regrouper plusieurs "mentions" d'une entité provenant d'un ou plusieurs textes. Nous détaillons dans la section 2.2.1 les méthodes employées pour la reconnaissance de ces entités nommées et la résolution de coréférence entre ces entités.

L'extraction de relations a pour objet d'étude les liens existants entre plusieurs entités : celle-ci peut-être binaire (entre deux objets) ou n-aire (plus de deux objets en relation). Il s'agit par exemple de détecter dans un corpus de documents que Barack Obama est l'actuel président des États-Unis, ce qui se traduira par une relation de type "président de" entre l'entité de type *Personne* "Barack Obama" et l'entité de type *Lieu* "États-Unis". La détection de relations n-aires correspond à ce que l'on nomme en anglais "record extraction" où il s'agit de repérer un réseau de relations entre entités et dont l'extraction

61. Traitement Automatique du Langage Naturel
62. Message Understanding Conference, http://www-nlpir.nist.gov/related_projects/muc/
63. Automatic Content Extraction, http://www.itl.nist.gov/iad/mig/tests/ace/

d'événements fait partie. Cette problématique diffère de la tâche de reconnaissance des entités nommées sur un point essentiel. Une entité nommée correspond à une portion séquentielle de texte et peut être directement représentée par une annotation délimitant le début et la fin de cette séquence. Une relation entre entités ne correspond pas directement à un ensemble consécutif de mots mais représente un lien entre deux portions de texte. Cela implique des processus et des formats d'annotation différents pour ces deux types de tâches. Dans la section 2.2.2, nous nous focaliserons sur les méthodes utilisées pour la détection de relations binaires, les relations n-aires seront abordées par le biais de l'extraction des événements. Est couramment associée à cette tâche d'EI l'extraction d'attributs ayant pour objectif d'extraire automatiquement un ensemble pré-défini de propriétés rattachées à ces entités, le type de ces propriétés dépendant directement de la nature de l'objet en question. Par exemple pour une personne, on pourra extraire son nom et son prénom, sa nationalité, son âge, etc. et pour une entreprise, son siège social, le nombre d'employés, etc.

L'extraction des événements est une autre tâche de l'EI très étudiée. 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. Comme cela a été décrit dans les sections 1.3 et 1.3.1.2, cette définition peut varier selon les points de vue théoriques et les applications et donne lieu à différentes représentations et ontologies dédiées aux événements. La détection d'événements s'avère particulièrement utile dans les activités de veille en général et intéresse de plus en plus les entreprises de nombreux domaines pour ses applications en intelligence économique et stratégique [Capet et al., 2011]. Nous présentons par la suite (en section 2.2.3) un tour d'horizon des techniques utilisées pour le repérage des événements dans les textes.

2.2 Approches d'extraction

En extraction d'information émergent historiquement deux principaux types d'approche : l'extraction basée sur des techniques linguistiques d'un côté et les systèmes statistiques à base d'apprentissage de l'autre. Cette distinction se retrouve largement en IA et dans les autres disciplines du TALN telles que la traduction automatique, etc.

Le premier type d'approche est nommé dans la littérature "approche symbolique", "à base de connaissances", "déclarative", "à base de règles" ou encore "système-expert" et exploite les avancées en TALN. Celles-ci reposent principalement sur une définition manuelle de toutes les formes linguistiques permettant d'exprimer l'information ciblée, autrement dit l'ensemble des contextes d'apparition de telle entité ou relation. Cela se traduit généralement par l'utilisation de grammaires formelles constituées de règles et patrons linguistiques élaborés par des experts-linguistes. Ces patrons sont généralement de deux types : les patrons lexico-syntaxiques et les patrons lexico-sémantiques. Les premiers associent des caractéristiques de mot (forme fléchie, lemme, genre, casse, etc.) à des indices structurels et de dépendance (syntagmatiques, phrastiques ou textuels). Les patrons lexico-sémantiques y ajoutent des éléments sémantiques tels que la projection de lexiques (*gazetteers*), l'utilisation de réseaux sémantiques externes tels que WordNet [64] ou encore d'ontologies [Hogenboom et al., 2011]. Les méthodes symboliques ont pour principales faiblesses leur taux de rappel peu élevé et leur coût de développement manuel coûteux.

64. http://globalwordnet.org

Le second type d'approche utilise des techniques statistiques pour apprendre des régularités sur de larges corpus de textes où les entités-cibles ont été préalablement annotées (domaine du "machine learning"). 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" (Hidden Markov Models, HMM), les "Champs Conditionnels Aléatoires" (Conditional Random Fields, CRF), les "Machines à Vecteur de Support" (Support Vector Machines, SVM), etc. ([Ireson et al., 2005] pour un état de l'art approfondi). Les principales limites de ce second type d'approche restent qu'elles nécessitent une grande quantité de données annotées pour leur phase d'apprentissage et produisent des modèles de type "boite noire" qui restent à l'heure actuelle difficilement accessibles et interprétables. Pour répondre à ce problème, des méthodes non-supervisées telles que [Etzioni et al., 2005] proposent d'utiliser des techniques de *clustering* pour extraire des entités d'intérêt.

Depuis quelques années, les méthodes hybrides tendent à se généraliser : les acteurs du domaine combinent plusieurs techniques face aux limites des approches symboliques et statistiques. 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 and Sekine, 2007], [Hobbs and Riloff, 2010]). Afin de diminuer l'effort de développement des systèmes symboliques, certains travaux s'intéressent à l'apprentissage automatique de règles d'extraction. A partir d'un corpus annoté, l'objectif est de trouver un ensemble minimal de règles permettant d'atteindre les meilleures performances, c'est-à-dire la meilleure balance entre précision et rappel (voir la section 2.5 pour une définition de ces mesures). Ces approches sont soit ascendantes ("bottom-up" en anglais) lorsqu'elles ont pour point de départ un ensemble de règles très spécifiques et opèrent des généralisations pour augmenter la couverture du système [Califf and Mooney, 2003] ; soit descendantes ("top-down" en anglais) lorsqu'elles partent d'un ensemble de règles très génériques pour arriver, par spécialisations successives, à un système plus précis [Soderland, 1999]. [Muslea, 1999] propose un tour d'horizon des grands types de règle résultant d'un apprentissage automatique.

Toutes ces méthodes reposent généralement sur des pré-traitements linguistiques dits "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). 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.

Nous détaillons par la suite quelques techniques couramment mises en œuvre selon le type d'objet à extraire : entité nommée, relation ou événement.

2.2.1 Extraction d'entités nommées et résolution de coréférence

La reconnaissance automatique d'entités nommées fut consacrée comme l'une des tâches principales de l'EI lors de la 6ème campagne d'évaluation MUC. Les systèmes alors proposés s'attellent à l'extraction de trois types d'entités définis sous les noms "ENAMEX", "TIMEX" et "NUMEX" correspondant respectivement aux noms propres (personnes, lieux, organisations, etc.), entités temporelles et entités numériques. Cette classification n'est pas la seule dans la littérature, [Daille et al., 2000] aborde notamment la distinction entre les catégorisations dites "référentielles" (telles que celle de MUC) et celles dites "graphiques". Les premières classent les entités nommées selon la nature des objets du monde auxquels elles renvoient tandis que les secondes proposent une classification selon la composition graphique des

entités. A titre d'exemple, une classification inspirée de [Jonasson, 1994] distingue les entités nommées "pures simples", des EN "pures complexes" et des EN "mixtes". Quelque soit la catégorisation choisie, celle-ci est fixée *a priori* dans la majorité des applications. A l'inverse, les récents travaux autour de la construction d'ontologies et de bases de connaissances à partir de textes mettent en œuvre une extraction "tous types confondus" pour en déduire *a posteriori* une classification.

L'extraction des entités nommées est généralement déroulée en deux phases : leur repérage dans le texte et leur typage selon la catégorisation pré-définie. [Nadeau and Sekine, 2007] propose un état de l'art des différentes approches explorées jusqu'à nos jours pour réaliser cette tâche : on y retrouve l'opposition "classique" (abordée dans la section précédente) entre méthodes symboliques et statistiques ainsi que l'intérêt récent pour les approches hybrides. Par ailleurs, [Friburger, 2006] présente les aspects linguistiques les plus communément exploités par les systèmes de REN.

Les approches à base de règles linguistiques sont les plus anciennes et, bien que coûteuses car elles sont définies manuellement par un expert, elles s'avèrent généralement plus rapides à l'exécution et plus personnalisables. L'ensemble de règles s'apparente à une grammaire locale et vise à définir de façon la plus exhaustive possible les différents contextes d'apparition de telle entité. De nombreux formats d'expression de règles d'extraction existent (CPSL [Appelt and Onyshkevych, 1998], JAPE [Cunningham et al., 2000], DataLog [Ceri et al., 1989], etc.) et celles-ci partagent la structure générique suivante :

$$regle : contexte \rightarrow action$$

La partie $contexte$ est composée de plusieurs propositions décrivant un contexte textuel au moyen de diverses caractéristiques linguistiques et formelles. Le repérage des entités nommées se base généralement sur la présence de majuscules puis leur catégorisation est réalisée grâce à des listes de noms propres connus ou de mots dits "catégorisants" (par exemple les prénoms). Ces caractéristiques peuvent être des attributs associés aux *tokens* (unités lexicales correspondant plus ou moins à un découpage en mots), à des segments plus larges tels que les syntagmes, à la phrase ou au texte dans son entier. Elles peuvent provenir de l'entité elle-même ou de son co-texte (respectivement "internal evidence" et "external evidence" [McDonald, 1996]). Lorsque le *contexte* est repéré dans le corpus à traiter, la partie $action$ de la règle est exécutée. Il s'agit dans la plupart des cas d'apposer une annotation d'un certain type sur tout ou partie du contexte textuel repéré.

Afin de gérer d'éventuels conflits lors de l'exécution d'un ensemble de règles, la plupart des systèmes intègrent des polices d'exécution sous forme d'heuristiques (par exemple : la règle qui produit l'annotation la plus longue est privilégiée) ou d'attribution de priorité à chaque règle. Par ailleurs, les systèmes à base de règles sont couramment implémentés en cascade, c'est-à-dire que les annotations fournies par un premier ensemble de règles peuvent être réutilisées en entrée d'un second ensemble de règles. [Wakao et al., 1996] développe le système LaSIE, une chaine de traitement symbolique fondée sur des grammaires d'unification exprimées en Prolog. Cet extracteur d'EN pour l'anglais a été évalué sur un ensemble d'articles du Wall Street Journal et atteint une F-mesure d'environ 92%. A la même période, l'ancien Stanford Research Center propose FASTUS (sponsorisé par la DARPA), un automate à états finis non-déterministe pour l'extraction d'entités nommées (sur le modèle de MUC-4) dans des textes en anglais et en japonais [Hobbs et al., 1997]. Plus récemment, [Maurel et al., 2011] présente le système *open-source* CasEN dédié au traitement des textes en français et développé grâce au logiciel CasSys (fourni par la plateforme Unitex) facilitant la création de cascades de transducteurs. Cet outil a

notamment été testé lors de la campagne d'évaluation ESTER 2 [65] (voir la section 2.5) dont l'objectif est de comparer les performances des extracteurs d'EN sur des transcriptions de la parole.

Bien que les approches statistiques bénéficient d'un essor plus récent, de nombreuses techniques d'extraction ont été et sont encore explorées. Le principe sous-jacent de ces méthodes est d'apprendre, à partir de textes pré-annotés avec les entités-cibles, un modèle de langage qui, appliqué sur un nouveau corpus, permettra d'extraire de nouvelles entités du même type. La majorité de ces approches se fonde sur un corpus d'apprentissage segmenté en *tokens* auxquels est associé un certain nombre de caractéristiques représentées par des vecteurs. Ces caractéristiques peuvent être intrinsèques au *token* (telles que sa forme de surface, sa longueur, sa catégorie grammaticale, etc.), liées à sa graphie (sa casse, la présence de caractères spécifiques, etc.) ou encore provenir de ressources externes (bases d'entités nommées connues, liste de mots catégorisants, etc.). A partir de ces différentes informations fournies dans le corpus d'apprentissage, la reconnaissance d'entités nommées est traitée comme un problème de classification. On distingue d'une part, des classifieurs dits linéaires tels que les modèles à base de régression logistique ou les machines à vecteurs de support [Isozaki and Kazawa, 2002] et, d'autre part, des classifieurs graphiques probabilistes tels que les HMMs [Zhao, 2004] ou les CRFs [Lafferty et al., 2001]. Le second type de classification est généralement plus performant car il permet de prendre en compte, lors de l'apprentissage du modèle d'annotation, les dépendances de classe entre *tokens* voisins. La principale limite de toutes ces approches reste le fait qu'elles construisent leur modèle au niveau *token* et ne permettent pas d'exploiter d'autres caractéristiques à un niveau de granularité plus élevé. Pour pallier à ce problème, il existe des techniques d'apprentissage statistique fondées sur un découpage en segments (en groupes syntaxiques, par exemple) ou sur une analyse de la structure globale des textes [Viola and Narasimhan, 2005]. Toutefois, les CRFs restent le modèle statistique le plus utilisé actuellement en REN de par leurs bonnes performances et leur capacité d'intégration de diverses caractéristiques [Tkachenko and Simanovsky, 2012].

Par ailleurs, les chercheurs en EI s'intéressent ces dernières années à combiner des techniques issues des approches symboliques et statistiques pour améliorer les performances de leurs systèmes d'extraction. [Charnois et al., 2009], par exemple, propose une approche semi-supervisée par apprentissage de patrons linguistiques pour l'extraction d'entités biomédicales (ici les noms de gènes). Ce travail met en œuvre une extraction de motifs séquentiels fréquents par fouille de textes et sous contraintes. Cette approche permet une extraction plus performante en utilisant un nouveau type de motif appelé LSR (utilisation du contexte du motif pour augmenter sa précision). Par ailleurs, [Charton et al., 2011] s'intéresse à l'extraction d'entités nommées en utilisant des motifs d'extraction extraits à partir de Wikipedia pour compléter un système d'apprentissage statistique par CRF. Les auteurs exploitent ici le contenu riche en noms propres et leurs variantes des ressources encyclopédiques telles que Wikipedia pour en extraire des patrons linguistiques. Ces résultats sont ensuite fusionnés avec ceux de l'approche statistique et une amélioration de la REN est constatée après évaluation sur le corpus de la campagne ESTER 2. L'inconvénient ici étant le besoin de données annotées, d'autres méthodes proposent d'interagir avec l'utilisateur : celui-ci fournit un petit ensemble d'exemples permettant d'obtenir un premier jeu de règles et celui-ci est amélioré de façon interactive et itérative [Ciravegna, 2001]. D'autres travaux tels que [Mikheev et al., 1999] choisissent de limiter la taille des *gazetteers* utilisés pour la REN et montrent que, tout en allégeant fortement la phase de développement de leur système (une approche hybride combinant un système à base de règles à un modèle probabiliste à Maximum d'Entropie), cela à un impact faible sur ses performances. [Fourour, 2002] propose Nemesis, un outil de REN fondé sur une approche incrémentielle se

65. Évaluation des Systèmes de Transcription Enrichie d'Émissions Radiophoniques, http://www.afcp-parole.org/camp_eval_systemes_transcription/

déroulant en trois phases : un premier jeu de règles est appliqué, celui-ci est suivi d'une étape d'apprentissage pour améliorer ce jeu, puis un second repérage est effectué grâce au jeu de règles amélioré.

Pour finir, beaucoup de travaux se penchent sur la résolution de coréférence entre entités nommées. En effet, comme introduit dans la section 2.1, il est courant de faire référence à une entité du monde réel (une personne, un lieu, une organisation, etc.) de plusieurs manières : par son/ses nom(s) propre(s) (par exemple, "Paris" ou "Paname"), par une expression la décrivant ("la capitale de la France") ou encore par un groupe nominal ou un pronom en contexte ("cette ville", "elle"). Ainsi, lorsqu'un texte contient différentes mentions d'une même entité, il apparait intéressant de pouvoir les lier (on obtient ainsi des chaines de référence) pour indiquer qu'elles font référence à un seul et même objet réel, qu'il y a donc coréférence. Cette problématique n'est pas triviale et peut s'appliquer à un seul et même document mais aussi au sein d'un corpus (coréférence entre entités provenant de textes distincts). La résolution de coréférence peut se faire de façon endogène ou exogène (en utilisant des ressources externes), en utilisant des techniques linguistiques, statistiques ou hybrides. Cette tâche fait notamment partie de campagnes d'évaluation telles que ACE (*Global Entity Detection and Recognition*). [Bontcheva et al., 2002] adopte une approche à base de règles pour la résolution de coréférence entre ENs et pronominale. Ces travaux ont été implémentés sous la forme d'un transducteur à états finis au sein de la plateforme GATE et sont intégrés à la chaine de traitement ANNIE (il s'agit des modules nommés Orthomatcher et Pronominal Corerefencer). Un exemple d'approche statistique et exogène est celle présentée par [Finin et al., 2009] où la base de connaissances Wikitology (construites à partir de Wikipedia, DBPedia et FreeBase) est utilisée en entrée d'un classifeur SVM (avec une trentaine d'autres caractéristiques textuelles) pour la construction de chaines de référence.

2.2.2 Extraction de relations

L'extraction de relations consiste à détecter et typer un lien exprimé textuellement entre deux entités. Cette tâche a notamment été proposée à l'évaluation lors de la campagne MUC-7, évaluation poursuivie par les campagnes ACE à partir de 2002. Les avancées dans cette problématique proviennent essentiellement des travaux d'EI menés dans les domaines de la médecine et de la biologie, notamment pour l'analyse des rapports médicaux et expérimentaux.

Les premiers systèmes développés sont fondés sur un ensemble de règles d'extraction dont les plus simples définissent des patrons sous forme de triplets du type : e_1 *relation* e_2 (où e_1 et e_2 sont deux entités reconnues au préalable). La majorité des relations n'étant pas exprimées aussi simplement dans les textes réels, les règles d'extraction doivent être plus élaborées et intégrer d'autres caractéristiques textuelles situées soit entre les deux entités visées soit autour du triplet relationnel. Comme pour la REN, les systèmes complètent couramment des caractéristiques de mot et une analyse structurelle par des indices sémantiques. [Muller and Tannier, 2004] présente une méthode symbolique comprenant une analyse syntaxique pour la détection de relations temporelles entre entités de type "événement". Dans le domaine médical, [Fundel et al., 2007] développe RelEx, un outil visant à extraire les interactions entre protéines et gènes et dont l'évaluation (sur le corpus MEDLINE) a montré de bonnes performances (précision et rappel d'environ 80%). [Nakamura-Delloye and Villemonte De La Clergerie, 2010] propose une méthode d'extraction de relations entre entités nommées basée sur une analyse en dépendance. Leur idée est de repérer les chemins syntaxiques entre deux entités (i.e. l'ensemble des relations de dépendance qu'il faut parcourir pour relier ces deux entités) afin de construire par généralisation des groupes de patrons de relations syntaxiques spécifiques à tel type de relation sémantique.

De nombreux travaux se tournent également vers l'apprentissage automatique de patrons de relation afin de faciliter ou de remplacer le travail manuel de l'expert. [Cellier et al., 2010], par exemple, s'attache au problème de la détection et du typage des interactions entre gènes par apprentissage de règles linguistiques. Leur approche réutilise une technique employée à l'origine en fouille de données : l'extraction de motifs séquentiels fréquents. Celle-ci permet d'apprendre à partir d'un corpus annoté un ensemble de régularités pour les transformer, après validation par un expert, en règles d'extraction. Il faut noter ici que le corpus d'apprentissage n'est pas annoté avec les relations-cibles mais uniquement avec des caractéristiques de plus bas niveau (entités nommées de type "gène", catégories morpho-syntaxiques, etc.). De plus, l'ajout de contraintes permet de diminuer la quantité de motifs retournés par le système et ainsi faciliter le tri manuel fait par l'expert.

Du côté des approches statistiques, [Rosario and Hearst, 2004] s'intéresse à la détection de relations entre maladies et traitements (de sept types distincts dont "cures", "prevents", "is a side effect of", etc.) et compare plusieurs méthodes statistiques pour cette tâche. Une sous-partie du corpus MEDLINE 2011 est annotée manuellement par un expert du domaine afin d'entraîner et tester plusieurs modèles graphiques et un réseau de neurones. Ce dernier obtient les meilleures performances avec une précision d'environ 97%. Une autre approche statistique pour l'extraction de relations est celle de [Zhu et al., 2009] proposant de combiner un processus de "bootstrapping" et un réseau logique de Markov. Le premier permet d'initier l'apprentissage à partir d'un petit jeu de relations fournies par l'utilisateur et ainsi diminuer le besoin en données annotées. De plus, ce travail exploite les capacités d'inférence permises par les réseaux logiques de Markov afin d'augmenter les performances globales de leur système StatSnowball.

La plupart des approches supervisées étant dépendantes du domaine de leur corpus d'apprentissage, [Mintz et al., 2009] s'intéresse à une supervision dite "distante" en utilisant la base de connaissances sémantique Freebase [66]. Le principe de ce travail est de repérer dans un corpus de textes brut des paires d'entités étant en relation dans Freebase et d'apprendre des régularités à partir du contexte textuel de cette paire. L'apprentissage est implémenté ici sous la forme d'un classifieur à logique de régression multi-classes et prend en compte des caractéristiques lexicales (par exemple, la catégorie grammaticale des mots), syntaxiques (une analyse en dépendance) et sémantiques (un repérage des entités nommées).

Pour finir, nous pouvons citer quelques travaux comme ceux de [Hasegawa et al., 2004] ou [Wang et al., 2011] proposant d'appliquer des techniques de "clustering" à l'extraction de relations. Le premier décrit une méthode non-supervisée d'extraction et de catégorisation de relations entre EN par "clustering". Celui-ci s'opère par une première étape de représentation du contexte de chaque paire d'entités proches en vecteurs de caractéristiques textuelles. Puis, on calcule une similarité cosinus entre vecteurs qui est donnée en entrée d'un "clustering" hiérarchique à lien complet. On obtient ainsi un "cluster" par type de relation, relation nommée en prenant le mot ou groupe de mot le plus fréquent entre paires d'EN au sein du "cluster". D'autre part, [Wang et al., 2011] s'intéresse à la détection de relations en domaine ouvert et de façon non-supervisée. Pour cela, leur approche est d'extraire un ensemble de relations par plusieurs phases de filtrage puis de les regrouper par type en utilisant des techniques de "clustering". La première étape est réalisée par trois filtrages successifs : les phrases contenant deux ENs et au moins un verbe entre les deux sont sélectionnées, puis les phrases non-porteuses de relation sont évacuées par des heuristiques et enfin par un apprentissage à base de CRFs. Une fois l'ensemble des relations pertinentes extraites, celles-ci sont regroupées par type sémantique en utilisant un algorithme de "clustering de Markov". Cette méthode ne nécessite pas d'annoter les relations dans un corpus d'apprentissage, ni de fixer au préalable les différents types de relation à extraire.

66. http://www.freebase.com/

2.2.3 Extraction d'événements

L'extraction d'événements, parfois considérée comme une extraction de relations n-aires, consiste à repérer dans un ou plusieurs textes des événements d'intérêt tels que définis dans la section 1.3. Cette tâche peut se résumer par le fait de répondre à la question suivante : "Who did what to whom when and where ?". Certains modèles y ajoutent les questions "how ?" et "why ?". La littérature dans ce domaine montre que l'extraction des événements regroupe généralement plusieurs sous-tâches [Ahn, 2006] :

1. détection des marqueurs d'événement ;
2. affectation des attributs ;
3. identification des arguments ;
4. estimation des rôles ;
5. résolution de coréférence.

Plusieurs campagnes MUC [67] 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.

La première approche symbolique retenue est décrite dans [Aone and Ramos-Santacruz, 2000] : il s'agit du système REES [68] permettant l'extraction de relations et d'événements à grande échelle. Cet outil repose sur l'utilisation combinée de lexiques et de patrons syntaxiques pour la détection d'événements principalement basés sur des verbes. Ces lexiques correspondent à une description syntaxique et sémantique des arguments de chaque verbe déclencheurs d'événement. Ces informations sont par la suite réutilisées au sein des patrons syntaxiques décrivant les différents contextes d'apparition d'un événement. Dans la lignée, [Grishman et al., 2002b] s'intéresse à la détection d'événements épidémiques au moyen d'un transducteur à états finis. Par ailleurs, le système d'extraction IE^2 (*Information Extraction Engine*) a été développé par la société américaine SRA International spécialisée dans le traitement de l'information [Aone et al., 1998]. Cet outil a obtenu les meilleures performances (51% de F-mesure) pour la tâche *Scenario Template* (ST) lors de la campagne d'évaluation MUC-7. Il s'agit d'une extracteur modulaire comprenant 6 modules dont celui nommé "EventTag" permettant le remplissage de scénarios d'événement grâce à des règles syntactico-sémantiques élaborées manuellement. Un autre outil d'EI à base de connaissances est celui proposé par [Appelt et al., 1995], il s'agit du système FASTUS ayant participé à plusieurs campagnes d'évaluation (MUC-4, MUC-5 et MUC-6). Cet extracteur est fondé sur un ensemble de règles de grammaire développé par des experts et a montré de bonnes performances dans la tâche d'extraction d'événements de MUC-6 [69]. FASTUS a obtenu une F-mesure de 51% contre 56% pour le meilleur des systèmes de la campagne. Cet outil est fondé sur un formalisme d'expression de règles nommé FASTSPEC visant à faciliter l'adaptation de l'outil à de nouveaux domaines.

Du côté des approches statistiques, [Chieu, 2003] développe le système ALICE [70] afin d'extraire des événements par apprentissage statistique. Ceux-ci ont évalué quatre algorithmes de classification issus de la suite Weka sur les données-test de la campagne MUC-4. Le corpus d'apprentissage est constitué des documents sources (des dépêches de presse) et des fiches d'événements associés. Les caractéristiques utilisées intègrent notamment une analyse des dépendances syntaxiques par phrase du corpus ainsi que des

67. Message Understanding Conference
68. Relation and Event Extraction System
69. la tâche concernait les changements de personnel dans la direction d'une entreprise
70. Automated Learning-based Information Content Extraction

chaines de coréférence entre entités nommées. Les meilleurs résultats sont obtenus avec un classifieur à Maximum d'Entropie (ALICE-ME) et celui-ci approche les performances du meilleur des participants de la campagne MUC-4. D'autre part, [Jean-Louis et al., 2012] présente un outil d'extraction d'événements sismiques exploitant des techniques d'analyse du discours et d'analyse de graphes. La reconnaissances des événements s'effectue en trois phases : un découpage des textes selon leur contenu événementiel, la construction d'un graphe des entités reconnues et le remplissage d'un formulaire d'événement par sélection des entités pertinentes dans le graphe. La première phase repose sur un apprentissage statistique par CRF tandis que la seconde est réalisée grâce à un classifieur à Maximum d'Entropie. Le corpus d'apprentissage pour ces deux premières étapes est constitué de dépêches provenant de l'AFP et de Google Actualités pour lesquelles des experts ont manuellement construits les formulaires d'événements. Les caractéristiques pour l'apprentissage (découpage en mots et phrases, détection des ENs, analyse syntaxique, etc.) ont été obtenues de l'analyseur LIMA [Besançon et al., 2010]. Enfin, le remplissage des formulaires d'événement est réalisé par combinaison de plusieurs algorithmes de sélection (PageRank, vote, etc.) afin de choisir la meilleure entité du graphe pour chaque champ du formulaire.

Les méthodes d'apprentissage de patrons ou les approches semi-supervisées apparaissent intéressantes comme par exemple le système de [Xu et al., 2006]. Ceux-ci proposent un outil d'extraction de patrons linguistiques par une méthode de "bootstrapping" appliquée à la détection des événements comme des remises de prix ("prize award events"). Cette approche est itérative et faiblement supervisée car elle permet, en partant de quelques exemples d'événements provenant d'une base de données existante, d'apprendre des régularités d'occurrence de ces événements et d'en déduire des patrons d'extraction. Ceux-ci ont ensuite été implémentés sous forme de règles dans une application créée grâce à la plateforme SProUT [71] [Drozdzynski et al., 2004]. Nous pouvons égelement citer le projet TARSQI [72] (respectant la spécification TimeML) qui a donné lieu au développement du système Evita [73]. [Saurí et al., 2005] présente succinctement les principes théoriques sur lesquels repose cet outil ainsi que son fonctionnement général. Les auteurs définissent les verbes, noms et adjectifs comme les trois catégories de mots déclencheurs étant les plus porteuses de sens pour la détection d'événements. Ils détaillent par la suite les différentes méthodes d'extraction associées à chaque type de déclencheur et plus particulièrement les caractéristiques textuelles et grammaticales à prendre en compte. Ainsi, pour la détection d'événements portés par un verbe, Evita opère un découpage en syntagmes verbaux et détermine pour chacun sa tête ; puis, vient une phase de tri lexical pour écarter les têtes ne dénotant pas un événement (verbes d'état, etc.) ; l'on tient ensuite compte des traits grammaticaux du verbe tels que la voix, la polarité (positif/négatif), la modalité, etc. ; et une analyse syntaxique de surface vient aider à l'identification des différents participants de l'événement. Pour finir, [Huffman, 1995] propose LIEP [74], un système de découverte de patrons d'extraction dont les résultats sont utilisés par l'extracteur d'événements symbolique ODIE [75]. Dans cette approche, on propose à l'utilisateur une interface lui permettant de remplir une fiche d'événement correspondant à une phrase donnée. Ces éléments sont ensuite utilisés pour apprendre par une approche ascendante (voir section 2.2) un ensemble de patrons récurrents qui sont ensuite ajoutés en tant que nouveaux chemins dans le transducteur à états finis de l'outil ODIE. Les auteurs de ce système montrent que LIEP approche les performances d'un système purement symbolique avec une F-mesure de 85% contre 89% pour ODIE.

71. Shallow Processing with Unification and Typed feature structures
72. Temporal Awareness and Reasoning Systems for Question Interpretation
73. Events In Texts Analizer
74. Learning Information Extraction Patterns
75. On-Demand Information Extraction

2.3 Plateformes et logiciels pour l'EI

Ces années de recherche en EI ont donné lieu comme vu précédemment à de nombreux travaux et, par conséquent, au développement de nombreux outils d'extraction d'information. Afin de faciliter ces développements et leur réutilisation au sein de chaines plus complexes, est apparu un nombre important de ce qu'on pourrait appeler des boites à outils pour l'EI et le TAL plus généralement. Celles-ci partagent une même visée finale, à savoir le traitement automatique des textes, mais se différencient sur plusieurs points. Elles proviennent tout d'abord de milieux différents, soit académique, soit industriel, soit sont issues d'une collaboration entre laboratoire(s) de recherche et entreprise(s) au sein d'un projet commun. De plus, il peut s'agir soit de simples entrepôts d'outils et algorithmes, soit de véritables plateformes d'intégration de modules hétérogènes. Dans ce cas, on pourra constater des choix d'architecture distincts pour la combinaison et l'enchainement de ces différents modules. Par ailleurs, la diversité et la complexité des documents traités constitue également un facteur de variation (différents formats, langues, domaines, structures, etc.). Cette hétérogénéité des contenus se traduit naturellement par différents choix de représentation de l'information, même si le format XML tend à se généraliser. Enfin, ces boites à outils ne donnent pas la même priorité à l'interaction avec l'utilisateur et ne se dotent pas des mêmes moyens de visualisation [Enjalbert, 2008].

Nous présentons ici un rapide tour d'horizon de ces boites à outils, centré sur différentes plateformes et suites logicielles distribuées en *open-source* et/ou gratuitement. Précisons tout d'abord que cette liste est non-exhaustive et a été constituée, sans ordre préférentiel, au fil de notre état de l'art et de nos travaux.

OpenCalais

Tout d'abord, la société Thomson Reuters (qui a racheté ClearForest) propose plusieurs services autour de l'extraction d'information regroupés sous le nom OpenCalais [76]. Celle-ci a mis en place OpenCalais Web Service, un outil en ligne d'extraction d'entités nommées, relations et évènements. Cet outil ainsi que les divers *plugins* qui l'accompagnent (Marmoset, Tagaroo, Gnosis, etc.) sont utilisables gratuitement pour usage commercial ou non. Le service d'annotation en ligne permet de traiter des textes en anglais, français et espagnol grâce à une détection automatique de la langue du texte fourni. Il extrait pour toutes ces langues un nombre conséquent de types d'entités nommées (villes, organisations, monnaies, personnes, e-mails, etc.) et attribue également un indice de pertinence/intérêt à chacune des extractions. L'analyse des textes en anglais est plus complète : extraction d'événements et de relations, désambiguïsation d'entités, détection de thème, association automatique de mots-clés ("semantic tags"), etc. Toutes ces annotations peuvent être récupérées au format RDF [77]. Enfin, précisons qu'OpenCalais fournit aussi bien des modules fondés sur des techniques linguistiques que des méthodes statistiques ou hybrides.

76. http://www.opencalais.com/
77. Ressource Description Framework

LingPipe

LingPipe développé par Alias-i [78] constitue une autre véritable "boîte à outils" pour l'analyse automatique de textes. Divers outils y sont disponibles gratuitement pour la recherche et, parmi ceux-ci, une majorité relèvent plus ou moins directement du domaine de l'extraction d'information. D'une part, des modules de pré-traitement permettent de « préparer » le texte pour la phase d'extraction : analyse morpho-syntaxique, découpage en phrases, désambiguïsation sémantique de mots. D'autre part, LingPipe met à disposition des modules de détection d'entités nommées, de phrases d'intérêt, d'analyse d'opinion et de classification thématique. Ces traitements sont tous réalisés par approche statistique et notamment par l'utilisation de CRF et d'autres modèles d'apprentissage (spécifiques à une langue, un genre de texte ou un type de corpus).

OpenNLP

Également reconnu, le groupe OpenNLP [79] rassemble un nombre important de projets *open-source* autour du Traitement Automatique du Langage. Son objectif principal est de promouvoir ces initiatives et de favoriser la communication entre acteurs du domaine pour une meilleure interopérabilité des systèmes. En extraction d'information, nous pouvons retenir les projets NLTK [80], MALLET [81], Weka [82] ou encore FreeLing [83]. Le premier correspond à plusieurs modules en Python pouvant servir de base au développement de son propre outil d'extraction. Le second projet, MALLET, est un logiciel développé par les étudiants de l'université du Massachussetts Amherst sous la direction d'Andrew McCallum, expert du domaine [McCallum, 2005]. Ce logiciel inclut différents outils pour l'annotation de segments (entités nommées et autres), tous basés sur des techniques statistiques de type CRF, HMM et MEMM [84]. Dans la même lignée, Weka est une suite de logiciels gratuits de "machine learning" développé à l'Université de Waikato (Nouvelle-Zélande) et distribués sous licence GNU GPL [85] [Hall et al., 2009]. Enfin, FreeLing [Padró and Stanilovsky, 2012] est une suite d'analyseurs de langage dont des modules de découpage en phrases, de tokenisation, lemmatisation, étiquetage grammatical, analyse syntaxique en dépendance, etc. Ce projet propose notamment une palette d'outils de traitement automatique pour la langue espagnole.

GATE

GATE est une plateforme *open-source* Java dédiée à l'ingénierie textuelle [Cunningham et al., 2002]. Créée il y a une vingtaine d'années par les chercheurs de l'université de Sheffield (Royaume-Uni), GATE est largement utilisé par les experts en TAL et dispose d'une grande communauté d'utilisateurs. Cela lui permet de disposer d'un ensemble de solutions d'aide et de support (forum, liste de diffusion, foire aux questions, wiki, tutoriels, etc.). Par ailleurs, les créateurs de GATE propose des formations ainsi que des certifications permettant de faire valoir ses compétences à l'utilisation de cette plateforme.

78. http://alias-i.com/lingpipe/
79. Open Natural Language Processing, http://opennlp.sourceforge.net/
80. Natural Language ToolKit, http://www.nltk.org/
81. Machine Learning for LanguagE Toolkit, http://mallet.cs.umass.edu/
82. Waikato Environment for Knowledge Analysis, http://www.cs.waikato.ac.nz/ml/weka/
83. http://nlp.lsi.upc.edu/freeling/
84. Maximum Entropy Markov Models
85. General Public License

LinguaStream

Par ailleurs, le GREYC [86] développe depuis 2001 la plateforme LinguaStream [87], un environnement intégré orienté vers la pratique expérimentale du TALN. Celui-ci permet un assemblage visuel de modules de traitement hétérogènes pour des applications en EI, RI, veille, résumé automatique, enseignement, etc. Ces enchainements de modules sont représentés sous forme de graphes acycliques et l'ensemble des données traitées est sérialisé en XML. [Widlocher et al., 2006] présente une application (réalisée dans le cadre de DEFT'2006) de segmentation thématique de textes implémentée grâce à la plateforme LinguaStream.

Unitex

La boite à outils Unitex est principalement développée par Sébastien Paumier à l'Institut Gaspard Monge (Université de Paris-Est Marne-la-Vallée) [Paumier, 2003]. Il s'agit d'une plateforme *open-source* multilingue pour le traitement en temps réel de grandes quantités de textes en langage naturel. Unitex permet d'appliquer des traitements divers sur les textes tels que le repérage de patrons linguistiques sous forme d'expressions régulières ou d'automates, l'application de lexiques et de tables, etc. Cela y est implémenté par des réseaux de transition récursifs définissables graphiquement. Cet outil propose également la production de concordances ou encore un ensemble d'études statistiques en corpus.

Nooj

Créateur de la plateforme Intex au LADL [88] (sous la direction du Professeur Maurice Gross), le Professeur Max Silberztein continue ses travaux depuis 2002 en proposant NooJ [89], un environnement de développement dédié au traitement du langage naturel écrit [Silberztein et al., 2012]. Cette suite propose des modules de traitement pour une vingtaine de langues (anglais, français, portugais, arabe, chinois, hébreu, etc.) et gère plus d'une centaine de formats de fichier d'entrée. Comme dans beaucoup de plateformes de ce type, les données y sont manipulées au format XML et enrichies grâce aux annotations fournies par les différents composants appliqués en cascade. Une communauté, en majorité européenne, s'est formée autour de cette plateforme donnant lieu depuis 2005 à une conférence NooJ annuelle réunissant divers travaux réalisés grâce à cet outil ainsi qu'à des tutoriels et ateliers réguliers.

Stanford NLP Group et Ontotext

Pour finir, mentionnons également les groupes de recherche Stanford NLP Group [90] et Ontotext [91] dont les travaux sont intégrés dans GATE. L'équipe de l'université de Stanford en Californie, a créé différents outils de TAL très utiles pour l'extraction d'information : un analyseur syntaxique probabiliste pour l'anglais, un étiqueteur morpho-syntaxique ainsi qu'un système d'extraction d'entités nommées qui

86. Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen
87. http://www.linguastream.org
88. Laboratoire d'Automatique Documentaire et Linguistique - Université de Paris-Est Marne-la-Vallée
89. http://www.nooj4nlp.net/
90. http://nlp.stanford.edu/
91. http://www.ontotext.com/

reconnaît les noms de personne, d'organisation et de lieu. Ontotext développe ses activités autour des technologies sémantiques et diffuse gratuitement la plateforme KIM [92] pour un usage non-commercial. Celle-ci propose de créer des liens sémantiques entre documents mais aussi d'extraire les entités nommées, relations et événements d'un texte et de les stocker automatiquement dans une base de données.

2.4 Applications

Les applications possibles de l'extraction automatique d'information sont à l'heure actuelle nombreuses et ne cessent de croître avec les avancées de la recherche (en particulier dans le domaine du Web). Dans cet ensemble d'applications, un petit nombre est historique et continue de susciter l'intérêt depuis les débuts de l'EI : il s'agit notamment de l'analyse des "news", du domaine biomédical ou encore de la veille économique et stratégique. D'autres usages sont plus récents et coïncident avec l'apparition de nouvelles technologies et des nouveaux besoins utilisateur ou techniques qui en découlent. Nous proposons ici un aperçu (non-exhaustif et général) de quelques cas d'application et des travaux de la littérature associés.

Tout d'abord, un grand nombre de travaux s'intéressent à l'utilisation des outils d'EI pour des besoins de veille : celle-ci peut être au service d'une entreprise, d'une entité gouvernementale ou encore d'un particulier souhaitant rester informé sur un sujet donné. Cette veille peut aider à la protection des populations par la prévention des épidémies ([Lejeune et al., 2010], [Grishman et al., 2002a], [Chaudet, 2004]) ou des événements sismiques [Besançon et al., 2011], par exemple. Par l'analyse automatique de différents types de sources d'information (rapports médicaux, dépêches de presse, réseaux sociaux, etc.), l'objectif est d'anticiper autant que possible ce genre de catastrophes et de suivre leur propagation pour assister les forces de secours par exemple. Les gouvernements portent aussi un grand intérêt à l'EI pour automatiser leurs processus de veille stratégique et militaire. [Zanasi, 2009], [Capet et al., 2008], [Tanev et al., 2008] ou encore [Pauna and Guillemin-Lanne, 2010] présentent leurs travaux pour une évaluation et un suivi du risque militaire et/ou civil, national et/ou international. [Hecking, 2003] se concentre sur l'analyse automatique des rapports écrits par les militaires, [Sun et al., 2005] et [Inyaem et al., 2010a] de leur côté s'intéressent à la prévention des actes de terrorisme. Enfin, [Goujon, 2002] présente une extraction automatique des événements appliquée à la crise de 2002 en Côte d'Ivoire. Dans le domaine économique, cette veille vise essentiellement à cerner des communautés de consommateurs (par exemple à partir du contenu des blogs [Chau and Xu, 2012]), à assurer un suivi des technologies et/ou produits d'un secteur donné pour les besoins d'une entreprise [Zhu and Porter, 2002] ou encore à améliorer son service-client par analyse des conversations téléphoniques [Jansche and Abney, 2002]. Pour tous les types de veille mis en œuvre, les acteurs du domaine s'intéressent également à adapter les processus d'EI pour garantir un traitement des informations en temps réel [Piskorski and Atkinson, 2011] [Liu et al., 2008]. Cette "fraicheur" des informations est particulièrement importante pour les analystes financiers et le suivi des évolutions des bourses par exemple [Borsje et al., 2010].

Par ailleurs, les chercheurs en EI se mettent au service d'autres sciences telles que la médecine et la biologie en permettant l'analyse automatique de rapports médicaux pour l'extraction d'interactions entre protéines, gènes, etc. [Rosario and Hearst, 2005] [Bundschus et al., 2008]. Dans un autre domaine d'intérêt public, la Commission Européenne a financé le projet PRONTO pour la détection d'événements dans les réseaux de transport public [Varjola and Löffler, 2010]. Les techniques d'EI sont également

92. Knowledge and Information Management

exploitées pour aider les acteurs judiciaires avec notamment l'analyse automatique des rapports de police [Chau et al., 2002]. Un autre cas d'application est le traitement automatique des publications scientifiques pour la construction de bases de citations telles que Citeseer [Lawrence et al., 1999]. De plus, comme mentionné en introduction de ce chapitre (voir la section 2.1), les résultats des outils d'extraction peuvent servir à d'autres disciplines de l'IA telles que la recherche d'information [Vlahovic, 2011], les systèmes de questions-réponses [Saurí et al., 2005], la construction des ontologies [Vargas-Vera and Celjuska, 2004] [Piskorski et al., 2007], le résumé automatique de documents [Radev et al., 2001], etc.

Enfin, avec la récente démocratisation de l'IA et l'entrée dans les foyers de nouvelles technologies de l'information, beaucoup de recherches en EI concernent des applications dédiées au grand public. Les systèmes de recommandation en ligne en bénéficient [Luberg et al., 2012] [Ittoo et al., 2006], mais aussi les logiciels anti-spam [Jason et al., 2004] ou encore les sites de recherche d'emploi [Califf and Mooney, 2003] [Ciravegna, 2001]. De plus, l'important volume de données récemment accessibles par le biais des réseaux sociaux, des sites de partage, collaboratifs, etc. a ouvert la voie à de nouvelles applications orientées vers le Web et en particulier le Web Sémantique. [Popescu et al., 2011] et [Sayyadi et al., 2009] s'intéressent à l'extraction d'événements dans les flux sociaux (Twitter notamment). [Nishihara et al., 2009] propose un système de détection des expériences personnelles dans les blogs. On peut également exploiter les méta-données fournies par les utilisateurs, comme le fait [Rattenbury et al., 2007] en analysant les "tags" postés sur Flickr pour en extraire des événements et des lieux dans les photos partagées. Pour finir, citons l'encyclopédie collaborative Wikipedia qui est non seulement une ressource précieuse pour le développement des outils d'EI mais qui bénéficie aussi de ces technologies [Chasin, 2010].

2.5 Évaluation des systèmes d'EI

2.5.1 Campagnes et projets d'évaluation

En parallèle des nombreux systèmes d'EI développés ces dernières années (dont certains ont été présentés dans les sections précédentes), la communauté des chercheurs a mis en place un certain nombre de campagnes d'évaluation telles que ACE, MUC, ESTER, CONLL [93], TAC [94], etc.

Afin de stimuler le développement des techniques d'extraction d'information et de dégager les pistes de recherche les plus prometteuses, ces campagnes d'évaluation sont menées tant au niveau national qu'international. Celles-ci ont pour but de mettre en place un protocole d'évaluation commun permettant aux experts du domaine de mesurer les performances de leurs outils. Les campagnes définissent généralement plusieurs tâches à accomplir telles que l'extraction d'entités nommées, de relations ou encore d'événements, la résolution de coréférence, etc. Le protocole le plus courant est de fournir un corpus d'entraînement et un corpus de test où les éléments à extraire ont été pré-annotés ainsi qu'un ou plusieurs scripts d'évaluation ("scoring"). Le corpus d'entraînement permet de préparer l'outil à la tâche d'extraction pour pouvoir ensuite s'auto-évaluer sur le corpus de test et estimer son score grâce aux scripts fournis. Une fois leurs systèmes préparés à la tâche d'évaluation, les participants sont évalués et classés par les organisateurs de la campagne. Ces évaluations s'accompagnent le plus souvent de publications d'articles dans lesquels ceux-ci décrivent leur outil et les techniques mises en œuvre. Cela

93. Conference on Computational Natural Language Learning, http://ifarm.nl/signll/conll/
94. Text Analysis Conference, http://www.nist.gov/tac/

permet de mettre en avant les nouvelles approches et de faire le point sur les performances de celles déjà connues.

Dans le domaine de l'extraction d'information, les campagnes MUC restent les pionnières et les plus connues au niveau international. Créées au début des années 1990 par la DARPA [95], elles constituent les premières initiatives pour encourager l'évaluation des systèmes d'extraction et ont fortement contribué à l'essor de ce domaine. À l'origine destinées au domaine militaire, les sept séries d'évaluation menées ont permis de diversifier les applications. Celles-ci se caractérisent par la tâche d'extraction consistant à remplir un formulaire à partir d'un ensemble de documents en langage naturel. Certains jeux de données de ces campagnes sont actuellement mis à disposition gratuitement.

La DARPA a également initié le « Machine Reading Program » (MRP) : projet visant à construire un système universel de lecture de texte capable d'extraire automatiquement la connaissance du langage naturel pour la transformer en représentation formelle. Celui-ci est destiné à faire le lien entre le savoir humain et les systèmes de raisonnement nécessitant ce savoir. Il s'agit pour cela de combiner les avancées en TALN et en IA.

Par ailleurs, nous pouvons citer le programme ACE (Automatic Content Extraction) qui, sous la direction du NIST [96], mène également des campagnes d'évaluation. Spécialisées dans l'analyse d'articles de presse, celles-ci évaluent l'extraction d'entités nommées et la résolution de co-référence (mentions d'entités nommées). Aujourd'hui, la campagne TAC (Text Analysis Conference) a pris la suite des actions menées dans le cadre du programme ACE.

Toujours à l'échelle mondiale, les campagnes CoNLL (Conference on Natural Language Learning) évaluent et font la promotion des méthodes d'extraction par apprentissage. Celles-ci sont classées parmi les meilleures conférences internationales dans le domaine de l'intelligence artificielle. Ce succès est en partie du au fait que ces conférences sont dirigées par l'ACL (Association of Computational Linguistics), la plus réputée des associations de linguistique et informatique. Celle-ci est aussi à l'origine des conférences Senseval/Semeval spécialisées dans l'évaluation des outils de désambiguïsation sémantique, point crucial en extraction d'information.

En Europe, l'association ELRA (European Language Ressources Association) a mis en place les conférences LREC (Language Ressources and Evaluation Conference). Lors de celles-ci les différents acteurs en ingénierie linguistique présentent de nouvelles méthodes d'évaluation ainsi que divers outils liés aux ressources linguistiques. De plus, cette association participe à l'évaluation de systèmes divers en fournissant les corpus et données nécessaires. Enfin, il nous faut citer la campagne française ESTER (Évaluation des Systèmes de Transcription Enrichie d'Émissions Radiophoniques) qui, entre autres activités, évalue le repérage d'entités nommées appliqué à des textes issus de transcription de la parole.

Les mesures d'évaluation les plus communément utilisées en extraction d'information sont la précision, le rappel et la F-mesure. Ces métriques peuvent être définies ainsi :

$$\text{Précision}_i = \frac{\text{nombre d'entités correctement étiquetées } i}{\text{nombre d'entités étiquetées } i}$$

95. Defense Advanced Research Projects Agency
96. National Institute of Standards and Technology

$$\text{Rappel}_i = \frac{\text{nombre d'entités correctement étiquetées } i}{\text{nombre d'entités } i}$$

$$F\text{-mesure}_i = \frac{(1+\beta^2)\cdot(\text{précision}_i \cdot \text{rappel}_i)}{(\beta^2 \cdot \text{précision}_i + \text{rappel}_i)}$$

où $\beta \in \mathbb{R}_+$

β est une facteur de pondération permettant de favoriser soit la précision soit le rappel lors du calcul de la F-mesure. La plupart des travaux pondère de façon égale la précision et le rappel, il est donc plus fréquemment utilisé une F_1-mesure définie comme suit :

$$F_1\text{-mesure}_i = \frac{2 \cdot \text{précision}_i \cdot \text{rappel}_i}{\text{précision}_i + \text{rappel}_i}$$

Précisons également que ces métriques sont souvent évaluées lors de la conception des approches afin de favoriser soit la précision soit le rappel de celles-ci en fonction de l'application visée.

2.5.2 Performances, atouts et faiblesses des méthodes existantes

Bien que le système d'évaluation actuel ne soit pas parfait, les différentes campagnes d'évaluation menées depuis plus de vingt ans permettent de dresser un bilan des performances des outils développés et de comparer les atouts et faiblesses des différentes approches adoptées. Comme nous l'avons exprimé précédemment, les avancées en EI sont disparates, elles varient en fonction de nombreux paramètres tels que l'objet ciblé (sa nature et sa complexité intrinsèque), l'ancienneté de la tâche en EI, le domaine/genre des textes analysés, les techniques employées, etc. Tout cela, rend très difficile une comparaison quantitative des systèmes développés et un bilan qualitatif nous parait plus approprié. Pour ce faire, nous nous inspirons de [Hogenboom et al., 2011] qui propose une évaluation selon quatre critères : la quantité de données annotées nécessaire, le besoin de connaissances externes, la nécessité d'une expertise humaine et l'interprétabilité des résultats. Nous donnons tout de même, à titre indicatif, quelques résultats (en termes de précision, rappel et F-mesure) issus des campagnes d'évaluation présentées ci-dessus.

En premier lieu, les systèmes purement linguistiques, bien que très précis, ont pour principales faiblesses leur taux de rappel moindre et leur coût de développement manuel coûteux. Ceux-ci ne nécessitent pas de corpus annoté mais impliquent un fort besoin en expertise humaine et dépendent souvent de connaissances externes (listes de mots, réseaux lexicaux, bases de connaissances, etc.). Ce dernier point se vérifie particulièrement dans le cas des systèmes à base de règles lexico-sémantiques. Un avantage non-négligeable de ces approches reste leur caractère symbolique permettant, sous réserve d'une expertise en TAL, d'appréhender plus facilement leur machinerie interne, de les adapter après analyse des résultats et d'observer dans la foulée l'impact des modifications. Ce cycle d'ingénierie est devenu plus aisé avec l'apparition des boites à outils pour l'EI (voir la section 2.3) dont certaines proposent des modules d'évaluation en temps réel de la chaine d'extraction. A titre d'exemple, le meilleur participant de la dernière campagne MUC pour l'extraction des événements (tâche *Scenario Template* de MUC-7) est un système symbolique et obtient les scores suivants : 65% de précision, 42% de rappel et 51% de F-mesure. Sur la tâche de REN, [Appelt, 1999] rapporte un taux d'erreur de 30% inférieur pour les approches symboliques comparées aux méthodes statistiques entièrement supervisées (respectivement environ 96% et

93% de F-mesure). Les systèmes à base de connaissances sont également les plus performants pour la résolution de coréférence avec des résultats allant jusqu'à 63% de rappel et 72% de précision.

De leur côté, les approches statistiques permettent de couvrir de nombreux contextes d'apparition (leur rappel est généralement plus élevé) mais nécessitent une grande quantité de données annotées pour l'apprentissage du modèle sous-jacent. Cela constitue une réelle contrainte car les corpus d'apprentissage sont inégalement disponibles selon la langue ciblée, le domaine/genre des textes, etc. Quelques travaux présentés plus haut s'intéressent à cette problématique en diminuant la supervision nécessaire dans le développement de tels systèmes : c'est le cas du "clustering" ou des techniques de "bootstrapping". En contrepartie, l'apprentissage statistique nécessite peu ou pas d'expertise humaine et des ressources externes limitées. Toutefois, il reste l'inconvénient que ces approches produisent des modèles de type "boite noire" qui restent à l'heure actuelle difficilement accessibles et interprétables. Enfin, les méthodes statistiques ont montré leur efficacité tout particulièrement en contexte bruité, dans le traitement des transcriptions de l'oral, par exemple. La meilleure approche statistique de la campagne CoNLL obtient une F-mesure de 91% en reconnaissance d'entités nommées. Les méthodes d'extraction par apprentissage statistique testées lors du challenge PASCAL 2005 atteignent une F-mesure de 75% pour la tâche d'extraction de relations.

Pour finir, nous avons vu l'intérêt récent pour le développement d'approches hybrides afin d'exploiter les points forts des méthodes précédentes. Même si ce type d'approche n'est pas parvenu pour le moment à éviter tous les écueils pointés ci-dessus, la combinaison des techniques symboliques et statistiques présente plusieurs atouts. L'apprentissage symbolique permet, par exemple, de diminuer l'effort de développement des règles d'un système-expert classique tout en augmentant le rappel de ces approches. On peut également opter pour une construction automatique des ressources linguistiques externes dont dépendent beaucoup des outils développés. Après analyse des erreurs d'extraction par méthode statistique, il est aussi intéressant de compléter le système par un ensemble de règles pour gérer les cas statistiquement peu fréquents. L'outil LP^2 (ayant remporté le challenge PASCAL 2005) implémente une méthode de déduction de règles d'extraction et obtient une F-mesure de près de 90% pour l'extraction de relations. Par ailleurs, dans le domaine médical, le système CRYSTAL montre une précision de 80% et un rappel de 75%.

Toutes approches confondues, les meilleurs systèmes en extraction de relations obtiennent une F-mesure d'environ 75% sur des données de la campagne ACE (ce score passe à 40% lorsque l'annotation des ENs est automatisée). Pour la tâche de remplissage de formulaires, les systèmes développés montent à une F-mesure de 60% (une annotation humaine obtenant 80%).

2.6 Problèmes ouverts

Pour conclure cet état de l'art sur l'extraction d'information, nous souhaitons faire le point sur les différents problèmes et challenges restant à résoudre. En effet, même si des progrès considérables ont été accomplis depuis les débuts de l'EI, un certain nombre de problèmes constituent toujours un réel frein à la commercialisation des systèmes existants [Piskorski and Yangarber, 2013].

Tout d'abord, la plupart des solutions sont développées pour un domaine ou un genre de texte particulier et voient leurs performances décroître rapidement face à des textes différents de ce point de vue. Le même problème survient lorsque les outils sont développés à partir de corpus très homogènes (sur la

forme ou le contenu) et que ceux-ci sont réutilisés sur d'autres corpus de nature plus variée. Ces limites concernent à la fois les méthodes symboliques et statistiques et nécessitent une ré-adaptation constante des techniques. Il a été montré que les performances d'un système d'extraction peuvent varier de 20% à 40% lors du passage à un nouveau domaine/genre de texte [Nadeau and Sekine, 2007]. Des travaux tels que [Chiticariu et al., 2010] ou [Daumé et al., 2010] s'attachent à faciliter la portabilité des outils d'EI d'un domaine à un autre tout en maintenant de bonnes performances.

Un autre enjeu est la réduction de l'effort de développement des systèmes, qu'ils soient symboliques ou statistiques. Outre leurs performances, ce point déterminera la commercialisation à grande échelle de ces outils. Du côté des approches à base de règles, cela est abordé par les travaux en apprentissage symbolique [Cellier and Charnois, 2010] tandis que pour réduire la quantité de données annotées nécessaires aux systèmes statistiques, les chercheurs se tournent vers des techniques comme l'apprentissage dit actif ("active learning") [Culotta et al., 2006] [Thompson et al., 1999]. Plus récemment, une solution prometteuse s'offre aux développeurs de systèmes supervisés et semi-supervisés, il s'agit du *crowdsourcing* : cette nouvelle pratique consiste à mettre à contribution les internautes pour créer du contenu et constitue un bon moyen pour la création de corpus annotés par exemple [Lofi et al., 2012].

Par ailleurs, on s'intéresse à adapter les méthodes d'extraction actuelles à une classification plus fine des entités nommées (villes, ONG, missiles, etc.) [Sekine et al., 2002] [Fleischman and Hovy, 2002]. En effet, la REN telle qu'elle était étudiée lors des premières campagnes d'évaluation atteint aujourd'hui des performances quasi-égales à celles d'un annotateur humain et ne répond que partiellement au besoin réel des utilisateurs finaux.

Au sujet de l'évaluation des technologies, nous pouvons souligner, d'une part, le peu de discussions dans la communauté de l'EI au sujet des métriques d'évaluation. En effet, ces métriques proviennent directement de l'évaluation en recherche d'information et s'avèrent, dans certains cas, peu adaptées à l'évaluation des systèmes d'extraction. [Lavelli et al., 2004] propose un résumé critique des méthodologies d'évaluation mises en œuvre depuis les débuts de l'EI. D'autre part, nous pouvons nous demander si les meilleurs résultats obtenus depuis quelques années en EI sont directement issus de l'amélioration des technologies ou s'il s'agit plutôt d'une simplification générale des tâches d'extraction.

Un dernier défi provient directement du récent engouement pour le Web Sémantique : il s'agit de tisser des liens entre les communautés de l'extraction d'information et de l'ingénierie des connaissances afin de "sémantiser" les informations extraites. Suivant l'objectif premier du Web Sémantique et du Web de données — favoriser l'émergence de nouvelles connaissances en liant les informations aux connaissances déjà présentes sur la toile — certains travaux s'attèlent à la problématique de création de liens sémantiques entre la sortie des extracteurs et les bases de connaissance existantes (notamment grâce au nommage unique par URI dans les données au format RDF). [Mihalcea and Csomai, 2007] [Ratinov et al., 2011] [Milne and Witten, 2008] sont des exemples de travaux de ce type dont le but est de lier les informations extraites à des concepts Wikipedia. Nous reviendrons plus amplement sur ces approches au chapitre 3.1.2.

2.7 Conclusions

La réalisation de cet état de l'art sur l'extraction d'information a révélé un domaine de recherche très étudié étant donné sa relative jeunesse : nous avons pu recenser un nombre important d'approches,

d'applications possibles, de logiciels et plateformes développés ainsi que de campagnes et projets d'évaluation menés jusqu'à nos jours. Les méthodes développées sont historiquement réparties en deux catégories : les symboliques et les statistiques. Les premières, développées manuellement par des experts de la langue, s'avèrent globalement plus précises, tandis que les secondes réalisent un apprentissage sur une grande quantité de données présentent généralement un fort taux de rappel. Parallèlement à cela, nous avons constaté une certaine complémentarité des approches existantes (voir la section 2.5.2) non seulement en termes de précision et rappel de façon générale mais également du point de vue des types d'entité ciblés, du genre textuel, du domaine d'application, etc. Il nous parait en conséquence pertinent de proposer un système d'extraction fondé sur la combinaison de plusieurs approches existantes afin de tirer partie de leurs différentes forces. Pour ce faire, les approches par apprentissage symbolique nous paraissent intéressantes car elles s'avèrent faiblement supervisées et plus flexible que d'autres approches statistiques. Enfin, ce tour d'horizon nous a permis de comparer différents outils et logiciels pour la mise en œuvre de ces approches ainsi que différents jeu de données potentiellement adaptés à l'évaluation de nos travaux.

Chapitre 3

Capitalisation des connaissances

Sommaire

Lorsque la quantité de documents disponibles dépasse un certain seuil, la problématique essentielle devient d'aider les analystes à analyser cette masse de données et à identifier les informations d'intérêt sans avoir à parcourir et synthétiser manuellement l'ensemble des documents. Dans ce contexte, les outils d'EI abordés en chapitre 2 se trouvent limités : en effet, la plupart réalise une analyse de l'information parcellaire, mono-document, mono-genre et mono-langue. De plus, comme nous l'avons montré, ces systèmes sont encore, à l'heure actuelle, imparfaits et, même si les progrès dans ce sens ont été significatifs depuis des années, les erreurs d'analyse ne sont pas rares. Face à cela, il devient de plus en plus nécessaire d'adopter un point de vue global et de concevoir un système de capitalisation des connaissances permettant à la fois d'extraire les informations d'intérêt à partir d'une masse de documents mais également de valoriser les résultats des extracteurs en assurant leur cohérence (éliminer les redondances, contradictions, créer des liens entres les informations, etc.) [Ji, 2010].

Cette problématique, relativement nouvelle, ne bénéficie pas encore d'un intérêt comparable aux deux premiers axes (chapitres 1 et 2) mais elle est l'objet de recherches dans divers domaines tels que la fusion de données, la réconciliation et le nettoyage de données, le *Linked Data*, la résolution de coréférence, la détection de similarité entre données, etc. Nous présentons dans ce chapitre quelques-unes des méthodes développées explorant différents angles d'un même axe de recherche, à savoir la capitalisation globale et automatisée des connaissances.

3.1 Fusion de données

La fusion de données a fait ses débuts aux États-Unis dans les années 70 et a été beaucoup étudiée jusqu'à nos jours dans des domaines divers tels que les applications militaires, de robotique, de transport ou encore en traitement d'images. L'objectif principal de cette discipline est d'optimiser l'acquisition de connaissances en combinant un ensemble d'informations (souvent imparfaites et hétérogènes) provenant de multiples sources plutôt que de les considérer chacune individuellement. Selon l'application, la fusion de données peut servir, d'une part, à reconstituer une situation le plus fidèlement possible à la réalité ou, d'autre part, à améliorer le processus de prise de décision [Desodt-Lebrun, 1996]. Parallèlement aux données à fusionner, la plupart des méthodes de fusion emploie des informations supplémentaires guidant la combinaison. Celles-ci peuvent provenir des données elles-mêmes ou de sources externes et sont par conséquent potentiellement exprimées dans des formalismes distincts. On distingue généralement plusieurs niveaux de fusion selon le type des informations traitées. Nous retiendrons la distinction principale entre la fusion dite numérique manipulant des informations de bas niveau (provenant essentiellement de capteurs) [Bloch, 2005] et la fusion dite symbolique dédiée aux informations de plus haut niveau. C'est dans le cadre de ce second type de fusion (symbolique) qu'apparaissent les travaux introduisant des techniques issues de l'intelligence artificielle et principalement ce que l'on nomme les systèmes à base de connaissance. Ceux-ci sont fondés, d'une part, sur une base de connaissances contenant l'expertise du domaine (les faits ou assertions connus) et, d'autre part, un moteur d'inférence permettant de déduire de nouvelles connaissances à partir de l'existant. Les systèmes à base de connaissance peuvent impliquer différents types de traitement sur les données : les raisonnements temporel et spatial, les déductions et inductions logiques, l'apprentissage automatique, diverses techniques de traitement automatique du langage, etc. La fusion symbolique est souvent choisie pour le traitement des données textuelles et trouve de nombreuses applications dans les systèmes automatiques nécessitant une interaction avec l'être humain.

Alors que beaucoup de recherches ont été menées dans le cadre de la fusion de données numériques, l'extraction automatique d'information apparait comme une perspective nouvelle permettant d'appliquer ces techniques à un autre type de données et dans un contexte particulièrement incertain et bruité. Ce besoin est fondé sur une constatation principale qui s'avère particulièrement vraie dans le contexte de la veille en sources ouvertes : la même information est rapportée de nombreuses fois par des sources différentes et sous diverses formes. Cela ce traduit notamment par l'utilisation de divers vocabulaires et conventions pour exprimer les mêmes données, des informations plus ou moins à jour et complètes selon les sources, des points de vue différents sur les faits, etc.

Dans les sections suivantes, nous abordons plusieurs axes de recherche traitant de cette problématique, à savoir la réconciliation des données, le Web de données et la détection de similarité entre données, puis nous nous centrerons sur l'application des méthodes de capitalisation de connaissances à la reconnaissance des événements.

3.1.1 Réconciliation de données

Le réconciliation de données est abordée dans la littérature sous de multiples dénominations, provenant de différentes communautés scientifiques et mettant l'accent sur un aspect particulier de cette problématique : des travaux comme ceux de [Winkler et al., 2006] parlent de *record linkage* (littéralement traduit par "liaison d'enregistrements ou d'entrées"), on trouve également les termes d'appariement d'objets (*object matching*), réconciliation de référence [Saïs et al., 2009], *duplicate record detection* (détection de doublons) [Elmagarmid et al., 2007], désambiguïsation d'entités ou encore résolution de co-référence entre entités [Bhattacharya and Getoor, 2007]. Les premiers sont plutôt issus de la communauté des bases de données, tandis que ces derniers sont généralement employés par les chercheurs en intelligence artificielle (TAL, ingénierie des connaissances, Web sémantique, etc.).

Le problème de la réconciliation de données a fait ses débuts dans les années 60 avec les travaux de [Newcombe et al., 1959] en génétique puis avec ceux de [Fellegi and Sunter, 1969] pour le traitement de duplicats dans des fichiers démographiques. La capacité à désambiguïser des dénominations polysémiques ou d'inférer que deux formes de surface distinctes réfèrent à la même entité est cruciale pour la gestion des bases de données et de connaissances. La méthode la plus simple (mais aussi la moins efficace) pour réconcilier des données est de comparer uniquement leurs représentations textuelles. Les premiers travaux dans ce sens réalisent une réconciliation de référence par paires de mentions et fondée sur la représentation de leurs contextes linguistiques en vecteurs de caractéristiques. Différentes mesures de similarité (voir la section 3.1.3) sont ensuite estimées entre ces vecteurs pour les combiner de façon linéaire par moyenne pondérée, par exemple [Dey et al., 1998]. Plus récemment, d'autres travaux proposent un appariement des descriptions de données plus générique mais toujours basé sur des comparaisons locales [Benjelloun et al., 2006] ou suivent une approche globale exploitant les dépendances existant entre les réconciliations [Dong et al., 2005]. Enfin, [Saïs et al., 2009] propose une approche de réconciliation de référence à base de connaissances et non-supervisée qui combine une méthode logique et une technique numérique. Ces travaux permettent d'exploiter la sémantique exprimée par les données et par leur structure par application d'un ensemble d'heuristiques.

Dans le domaine du traitement de données textuelles, lorsque cette tâche est réalisée sans liaison à une base de connaissances externe, elle est souvent appelée résolution de co-référence : les mentions d'entités provenant soit d'un même document soit de plusieurs sont regroupées, chaque groupe référant

à une seule et même entité réelle. La tâche de résolution de co-référence sur un ensemble de documents a été adressée par plusieurs chercheurs en commençant par [Bagga and Baldwin, 1999]. Ceux-ci se sont attelés au problème de la co-référence inter-documents en comparant, pour chaque paire d'entités dans deux documents distincts, les vecteurs de mots construits à partir de toutes les phrases contenant les mentions des deux entités. Pour aller plus loin, d'autres approches utilisent des modèles probabilistes [Verykios and Elmagarmid, 1999] mais ceux-ci nécessitent une grande quantité de données annotées pour leur apprentissage. [Wacholder et al., 1997] a développé *Nominator* l'un des premiers systèmes de REN et de co-référence entre entités fondé sur des mesures de similarité entre contextes d'occurrence. Par ailleurs, un ensemble de travaux récents proposent de représenter l'information extraite de plusieurs documents sous la forme d'un réseau [Ji et al., 2009] où les entités d'intérêt sont vues comme les feuilles d'un graphe et peuvent être liées entre elles par différents types de relations statiques.

Ces différentes études visent à regrouper toutes les mentions d'une même entité au sein d'une collection de textes. Toutefois, la construction d'une chaine de co-référence entre entités ne suffit pas à la capitalisation des connaissances car il reste nécessaire de rattacher cette chaine à une entité du monde réel. Il s'agit d'une tâche complexe et il s'avère parfois difficile, même pour un lecteur humain, de déterminer à quel objet il est fait référence dans un texte. Dans la plupart des cas, celui-ci identifie la référence d'une entité grâce à des indices issus de son contexte textuel mais aussi et surtout en exploitant l'ensemble du savoir qu'il a déjà acquis par des expériences passées.

3.1.2 Web de données

Avec l'essor du Web sémantique de nombreuses recherches proposent d'aller plus loin dans la réconciliation de données et de lier les connaissances entre elles pour créer un Web de données ou *Linked Open Data* (LOD). Ce liage d'entités est défini comme l'appariement d'une mention textuelle d'une entité et d'une entrée définie dans une base de connaissances.

On distingue 3 défis à dépasser pour cette tâche [Dredze et al., 2010] :

- Les variations de forme : une seule et même entité est souvent mentionnée sous diverses formes textuelles telles que des abréviations (*JFK* pour *John Fitzgerald Kennedy*), des expressions raccourcies (*Obama* pour *Barack Obama*), des orthographes alternatives (*Osama*, *Ussamah* ou encore *Oussama* pour désigner l'ancien dirigeant d'Al-Qaïda) et des alias/pseudonymes (*Big Appl*e pour la ville de *New York*).
- Les ambiguïtés référentielles : une seule et même forme de surface peut correspondre à plusieurs entrées dans une base de connaissances. En effet, de nombreux noms d'entités sont polysémiques (*Paris* est une ville en France mais aussi au Texas).
- L'absence de référent : il peut arriver, particulièrement lorsque la quantité de documents traités est conséquente, que la ou les base(s) de connaissances servant de référentiel ne contiennent pas d'entrée pour une ou plusieurs des entités repérées dans les textes (par exemple, le tout nouveau nom donné par les spécialistes à une catastrophe naturelle).

De par sa popularité et son exhaustivité, la base de connaissances Wikipédia a largement été utilisée comme base de référence pour le liage de données : ce cas particulier a même reçu le nom de *Wikification*. Etant donné une chaine de caractères identifiée dans un texte, l'objectif est de déterminer la page Wikipédia à laquelle cette chaine fait référence. Par exemple, pour la phrase suivante "Votre avion décollera de JFK", un système de Wification retournera l'identifiant `http://fr.wikipedia.org/wiki/`

A%C3%A9roport_international_John-F.-Kennedy, correspondant à l'aéroport de New York et non la page du 35ème président des États-Unis, par exemple. Les études existantes sur la Wikification diffèrent en fonction des types de corpus traités et des expressions qu'elles cherchent à lier. Par exemple, certains travaux se focalisent sur la Wikification des entités nommées alors que d'autres visent toutes les expressions d'intérêt en cherchant à reproduire un équivalent de la structure de liens de Wikipédia pour un ensemble de textes donné.

Le système *Wikifier* [Milne and Witten, 2008], par exemple, est fondé sur une approche utilisant les liens entre les articles de Wikipédia en tant que données d'apprentissage. En effet, les liens entre les pages de cette base étant créés manuellement par les éditeurs, ils constituent des données d'apprentissage très sûres pour réaliser des choix de désambiguïsation. Par ailleurs, le prototype LODifier [Augenstein et al., 2012] vise à convertir des textes en langage naturel de tout domaine en données liées. Cette approche incorpore plusieurs méthodes de TAL : les entités nommées sont repérées par un outil de REN, des relations normalisées sont extraites par une analyse sémantique profonde des textes et une méthode de désambiguisation sémantique (*Word Sense Disambiguation* en anglais) permet de traiter les cas de polysémie. L'outil Wikifier y est utilisé pour améliorer la couverture de la REN mais également pour obtenir des liens vers le Web de données grâce aux identifiants DBPedia proposés. Le sens d'un document est finalement consolidé en un graphe RDF dont les noeuds sont connectés à des bases à large couverture du LOD telles que DBPedia et WordNet. Contrairement aux approches précédentes, les travaux de [Dredze et al., 2010] sont facilement adaptables à d'autres bases de connaissances que Wikipédia. Cette approche implémente un apprentissage supervisé fondé sur un ensemble exhaustif de caractéristiques textuelles et s'avère particulièrement efficace dans les cas d'absence de référent.

Pour finir, citons l'initiative NLP2RDF [97] qui, également dans l'objectif de créer le Web de données, propose le format d'échange unifié NIF (NLP Interchange Format) afin de favoriser l'interopérabilité des méthodes et systèmes de TAL et l'exploitation de leurs résultats au sein du LOD (via le langage RDF notamment).

3.1.3 Similarité entre données

Comme entrevu dans les sections précédentes, les recherches menées en réconciliation de données (au sens large) vont de paire, dans la littérature, avec les travaux conduits autour des calculs de similarité entre ces données. La multitude des calculs de similarité existants faisant que nous ne pourrons les parcourir de façon exhaustive ici, nous choisissons de les présenter par catégories et ce en faisant référence à des états de l'art existants, spécialisés sur cette problématique.

Nous avons retenu deux d'entre eux très complets à savoir [Elmagarmid et al., 2007] et [Bilenko et al., 2003]. Les approches de calcul de similarité procèdent généralement en deux étapes : une phase dite de préparation des données suivie d'une phase de fusion des champs référant à une même entité. En effet, hétérogènes du point de vue de leur fond et de leur forme, les données manipulées nécessitent d'être pré-traitées dans le but de les stocker de façon la plus uniforme possible dans les bases de données. Cela consiste généralement à réduire au maximum leur diversité structurelle en les convertissant dans un format commun et normalisé. Vient, dans un second temps, l'étape de comparaison des données entre elles et d'estimation de leur similarité. Une grande quantité de méthodes ont été proposé pour ce faire : celles-ci varient sur plusieurs critères (type de données visé, niveau de comparaison, etc.) qui donnent

97. http://nlp2rdf.org/about

à des catégorisations différentes. [Elmagarmid et al., 2007] distingue les mesures de similarité au niveau caractère (distance d'édition, *affine gap distance*, distance de Smith-Waterman, distance de Jaro, Q-Grams) et les métriques basées sur les *tokens* (chaines atomiques, WHIRL, Q-Grams avec TF.IDF). [Bilenko et al., 2003] différencie les mesures statiques (distance d'édition, métrique de Jaro et ses variantes, distances basées sur les *tokens* et hybrides) de celles à base d'apprentissage (classifieurs SVM entrainés avec des vecteurs de caractéristiques, *affine gap distance*, modèles d'apprentissage avec distance de Levenstein). Pour finir, nous pouvons citer les recherches de [Moreau et al., 2008] définissant un modèle générique en vue de faciliter la combinaison de différentes mesures de similarité au sein d'un même système.

3.2 Capitalisation appliquée aux événements

Selon [Quine, 1985], deux mentions d'événement co-réfèrent si elles partagent les mêmes propriétés et participants. Contrairement à la co-référence entre entités dites simples, celle entre événements s'avère plus complexe principalement car les mentions d'événements présentent des structures linguistiques plus riches et variées que les mentions d'entités simples. De plus, le premier type de tâche est réalisé au niveau du mot ou du groupe de mots alors que la résolution de co-référence entre événements doit s'effectuer à un niveau plus élevé (phrase, discours).

Une partie des approches existantes pour répondre à ce problème repose sur des méthodes d'apprentissage supervisées explorant diverses caractéristiques linguistiques des textes [Humphreys et al., 1997] [Bagga and Baldwin, 1999] [Naughton et al., 2006]. [Lee et al., 2012], par exemple, propose une approche globale par apprentissage pour la résolution de co-référence, réalisée de façon conjointe entre entités et événements, au sein d'un seul ou de plusieurs documents. Celle-ci est fondée sur une méthode itérative de regroupement exploitant un modèle de régression linéaire appris sur ces données. Toutefois, la résolution de co-référence entre événements impliquant d'explorer une grande quantité de caractéristiques linguistiques, annoter un corpus d'apprentissage pour cette tâche requiert un effort de développement manuel important. De plus, étant donné que ces modèles reposent sur des décisions locales d'appariement, ils ne peuvent généralement pas capturer des relations de co-référence au niveau d'un sujet défini ou sur une collection de plusieurs documents. En réponse à cela, sont créés des systèmes comme *Resolver* [Yates and Etzioni, 2009] qui permet d'agréger des faits redondants extraits (par l'outil d'extraction d'information *TextRunner*) grâce à un modèle non-supervisé estimant la probabilité qu'une paire de mentions coréfèrent en fonction de leur contexte d'apparition (exprimé sous forme de n-tuples). Par ailleurs, [Chen and Ji, 2009] propose de représenter les co-références entre événements par un graphe pondéré non-orienté où les nœuds représentent les mentions d'événement et les poids des arêtes correspondent aux scores de co-référence entre deux des mentions. La résolution de co-référence est ensuite réalisée comme un problème de clustering spectral du graphe mais le problème le plus délicat reste l'estimation des similarités en elles-mêmes. Il nous faut noter enfin les travaux de [Khrouf and Troncy, 2012] explorant la problématique de la réconciliation des événements dans le Web de données. En effet, partant du constat que le nuage LOD contient un certain nombre de silos d'événements possédant leurs propres modèles de données, ceux-ci proposent d'aligner cet ensemble de descriptions d'événements grâce à diverses mesures de similarité et de les représenter avec un modèle commun (l'ontologie LODE présentée en section 1.3.1.2).

Pour finir, concernant l'évaluation de cette problématique, nous pouvons mentionner la campagne d'évaluation ACE (présentée en section 2.5.1) mettant à disposition des données d'évaluation pour la tâche *Event Detection and Recognition* (VDR). Toutefois, son utilisation s'avère limitée car cette ressource ne contient que des annotations de co-référence intra-documents et pour un nombre restreint de types d'événements (*Life*, *Movement*, *Transaction*, *Business*, *Conflict*, *Contact*, *Personnel* and *Justice*).

3.3 Conclusions

La réalisation de cet état de l'art a mis en exergue une suite logique à nos travaux sur l'extraction automatique d'information, à savoir la problématique du passage du texte à la connaissance proprement dite. Comme nous avons pu le voir, celle-ci a donné lieu à diverses recherches au sein de plusieurs communautés de l'IA, chacune d'elles manipulant sa propre terminologie adaptée à ses propres besoins. Ses divergences de vocabulaire n'empêchent pas de voir la place importante réservée à la capitalisation des connaissances au sein des recherches actuelles que ce soit en fusion de données, extraction d'information ou Web sémantique. Certains de ces travaux nous paraissent convenir à nos objectifs tels que [Chen and Ji, 2009] avec leur représentation en graphe de l'ensemble des connaissances (bien adaptée aux travaux dans le cadre du Web sémantique et du WebLab notamment), [Khrouf and Troncy, 2012] pour leur approche globale autour de plusieurs bases de connaissances et enfin, les différentes similarités entre données qui peuvent permettre de réconcilier des extractions. Les enseignements tirés de ce tour d'horizon ont été exploités lors de l'élaboration de notre approche d'agrégation des événements (voir le chapitre 6). Ce chapitre clôture notre partie état de l'art sur les différents domaines de recherche abordés par cette thèse.

Deuxième partie

Contributions de la thèse

Introduction

Cette seconde partie présente les contributions réalisées durant cette thèse en réponse à notre problématique de recherche et de façon adaptée aux conclusions de l'état de l'art réalisé. Nous y proposons, dans le chapitre *4 Modélisation des connaissances du domaine*, notre première contribution : un modèle de représentation des connaissances conçu en accord avec les besoins de notre cadre applicatif (le ROSO et le *Media mining*) et avec les observations faites au chapitre 1 (sur les modèles et approches existantes). Cette première proposition comprend, d'une part, une modélisation des événements en plusieurs dimensions et, d'autre part, une implémentation de ce modèle au sein d'une ontologie de domaine, nommée WOOKIE, élaborée durant nos recherches. Dans un second chapitre (*5 Extraction automatique des événements*), les contributions liées à notre axe de recherche Extraction d'information seront exposées. Nous commencerons par la conception d'une approche de reconnaissance d'entités nommées pour l'anglais et le français et implémentée grâce à la plateforme GATE. Puis, le cœur du chapitre sera dédié à l'élaboration d'une approche mixte pour l'extraction automatique des événements dans les textes selon le modèle de connaissances défini auparavant. Celle-ci est fondée sur deux techniques actuelles issue de la littérature en extraction d'information : une première méthode symbolique à base de règles linguistiques contextuelles et une seconde fondée sur un apprentissage de patrons d'extraction par fouille de motifs séquentiels fréquents. L'ensemble des méthodes exposées seront accompagnées d'exemples tirés de données réelles afin de faciliter leur compréhension. Enfin, le dernier chapitre de cette partie (*6 Agrégation sémantique des événements*) sera centré sur un processus d'agrégation sémantique des événements destiné à assurer la création d'un ensemble de connaissances cohérent et d'intérêt pour l'utilisateur. Cela sera réalisé en différentes phases (conformément aux observations faites durant l'état de l'art) : une première étape de normalisation des différentes extractions, suivie d'une approche permettant d'estimer une similarité sémantique multi-niveaux entre événements et un processus d'agrégation sémantique fondé sur une représentation en graphe des connaissances.

Chapitre 4

Modélisation des connaissances du domaine

Sommaire

4.1 Introduction

Ce chapitre détaille la première contribution proposée durant nos travaux de thèse : une modélisation des événements ainsi qu'une ontologie de domaine nommée WOOKIE. Celles-ci ont été élaborées en fonction des conclusions de notre état de l'art et de façon adaptée à notre problématique, à savoir l'extraction automatique des événements dans le cadre du ROSO. Nous proposons, tout d'abord, le modèle d'événement défini pour servir de guide à l'ensemble de notre processus de capitalisation des connaissances. Un événement est représenté selon quatre dimensions (conceptuelle, temporelle, spatiale et agentive) pour chacune desquelles nous avons défini des propriétés bien spécifiques. Enfin, nous présentons l'élaboration de notre ontologie de domaine au regard de la littérature et en y intégrant notre modélisation des événements. Nous terminons ce chapitre par un bilan des forces et faiblesses de notre contribution.

4.2 Notre modèle d'événement

Nous prenons pour point de départ la définition de Krieg-Planque ("un événement est une occurrence perçue comme signifiante dans un certain cadre") qui apparaît bien adaptée à nos travaux. Cette définition restant très théorique, il convient d'expliciter comment un événement est exprimé au sein des dépêches de presse (celles de l'AFP [98], par exemple). Après observation de plusieurs dépêches, celles-ci semblent généralement être centrées sur un événement principal, celui-ci étant le plus souvent résumé dans le titre et explicité tout au long de l'article (en faisant parfois référence à d'autres événements). Cette description de l'événement tout au long de la dépêche, est constituée de plusieurs sous-événements ("mentions d'événement" dans le modèle ACE) qui contribuent à la "mise en intrigue" mentionnée auparavant. Ces mentions d'événements sont généralement composées d'un terme déclencheur (dit aussi "nom d'événement" ou "ancre") associé à une ou plusieurs autres entités d'intérêt ("arguments" dans le modèle ACE) telles que des circonstants spatio-temporels (date et lieu de l'événement) et des participants (acteurs, auxiliaires, instruments, etc.). L'objectif de nos travaux est d'extraire automatiquement les mentions d'événement pertinentes pour notre application pour ensuite agréger celles qui réfèrent à un seul et même événement dans la réalité.

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 conséquent, 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 (que nous appellerons dimension conceptuelle), 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).*

98. Agence France Presse

Les sections suivantes décrivent comment chaque dimension de l'événement est modélisée.

Enfin, dans nos travaux, le "cadre" mentionné par Krieg-Planque est défini par l'ontologie de domaine WOOKIE [99] (voir la section 4.3) et plus précisément par la spécification de la classe événement ("Event") en différentes sous-classes et propriétés. Il détermine quels sont les entités et événements d'intérêt pour notre application.

4.2.1 La dimension conceptuelle

La dimension conceptuelle S d'un événement correspond au sens véhiculé par le nom porteur de cet événement. En effet, comme le souligne [Neveu and Quéré, 1996], l'interprétation d'un événement dépend étroitement de la sémantique exprimée par les termes employés pour nommer cet événement. Cette dimension équivaut à la propriété sémantique des événements évoquée par [Saval et al., 2009] et représente le type de l'événement, c'est-à-dire sa classe conceptuelle au sein de notre ontologie de domaine.

C'est la taxonomie des événements au sein de WOOKIE qui constitue le support principal de cette dimension conceptuelle. Nous avons défini pour notre application environ 20 types d'événement d'intérêt pour le renseignement militaire regroupés sous le concept de *MilitaryEvent*. Les différentes sous-classes d'événement sont les suivantes :

- *AttackEvent* : tout type d'attaque,
- *BombingEvent* : les attaques par explosifs,
- *ShootingEvent* : les attaques par armes à feu,
- *CrashEvent* : tous les types d'accidents,
- *DamageEvent* : tous les types de dommages matériels,
- *DeathEvent* : les décès humains,
- *FightingEvent* : les combats,
- *InjureEvent* : tout type d'événement entrainant des blessés,
- *KidnappingEvent* : les enlèvements de personnes,
- *MilitaryOperation* : tout type d'opération militaire,
- *ArrestOperation* : les arrestations,
- *HelpOperation* : les opérations d'aide et de secours,
- *PeaceKeepingOperation* : les opérations de maintien de la paix,
- *SearchOperation* : les opérations de recherche,
- *SurveillanceOperation* : les opérations de surveillance,
- *TrainingOperation* : les entrainements,
- *TroopMovementOperation* : les mouvements de troupes,
- *NuclearEvent* : tout type d'événement nucléaire,
- *TrafficEvent* : tout type de trafic illégal.

Cette taxonomie a été essentiellement constituée en nous inspirant des modélisations existantes dans le domaine telles que celles présentées en section 1.3.3. Mais également par observation des différents types d'événements rapportés dans des dépêches de presse sur des thèmes tels que les guerres en Afghanistan et en Irak ou encore les diverses attaques terroristes dans le monde.

99. Weblab Ontology for Open sources Knowledge and Intelligence Exploitation

4.2.2 La dimension temporelle

Pour la représentation des entités temporelles extraites nous utilisons le "Time Unit System" (TUS) proposé par [Ladkin, 1987]. Contrairement à la majorité des modèles théoriques dédiés à la logique temporelle [Fisher et al., 2005], ce formalisme s'avère plus applicable à des situations réelles de traitement des entités temporelles, notamment, par sa proximité avec les systèmes calendaires communément utilisés. 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*. [Ligeza and Bouzid, 2008] définit toutes les valeurs maximales pour le granule *jour* de la façon suivante :

- $max_3([g_1, 2]) = 29$ lorsque a_1 est une année bissextile
- $max_3([g_1, 2]) = 28$ lorsque a_1 est une année non-bissextile
- $max_3([g_1, g_2]) = 31$ quelque soit g_1 et lorsque $g_2 \in \{1, 3, 5, 7, 8, 10, 12\}$
- $max_3([g_1, g_2]) = 30$ quelque soit g_1 et lorsque $g_2 \in \{4, 6, 9, 11\}$

Une date est dite *légale* lorsqu'elle est valide au regard de cet opérateur et plus généralement du système calendaire courant et elle est dite *illégale* dans le cas contraire.

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).

De plus, 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. Expression temporelle complète : *T est complète lorsque* $\forall_{i \in \{1,2,3\}}, g_i \neq \emptyset$

Définition 3. Expression temporelle incomplète : *T est incomplète lorsque* $\exists_{i \in \{1,2,3\}}, g_i = \emptyset$

Nous listons ci-dessous toutes les formes possibles que peuvent revêtir les dates extraites une fois exprimées avec le formalisme TUS ainsi que des exemples :

- $[year, month, day]$, e.g. $[2011, 12, 14]$;
- $[year, month]$ e.g. $[2011, 12, \emptyset]$;
- $[month, day]$, e.g. $[\emptyset, 12, 14]$;
- $[year]$ e.g. $[2011, \emptyset, \emptyset]$;
- $[month]$ e.g. $[\emptyset, 12, \emptyset]$;

– $[day]$ e.g. $[\emptyset, \emptyset, 14]$.

Enfin, le modèle TUS introduit l'opérateur *convexify* permettant la représentation des intervalles temporels convexes. Prenant pour paramètres deux intervalles primaires i et j, $convexify(i, j)$ retourne le plus petit intervalle de temps contenant i et j. Par exemple, $convexify([2008], [2011])$ correspond à l'intervalle de 3 ans entre le premier jour de l'année 2008 et le dernier jour de 2011. Nous utilisons cet opérateur pour exprimer de façon unifiée les périodes de temps extraites des textes et permettre ainsi des calculs temporels grâce au modèle TUS. Par conséquent, une extraction telle que "from 2001 to 2005" est normalisé sous la forme $convexify([2001], [2005])$.

4.2.3 La dimension spatiale

Pour notre application, nous choisissons de représenter les entités spatiales comme des aires géographiques et d'utiliser les relations topologiques du modèle RCC-8 pour leur agrégation. En effet, ce modèle s'avère mieux adapté à la comparaison d'entités spatiales que ceux à base de points et fondés sur les coordonnées géographiques. 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 système. Dans le cadre du WebLab, nous nous intéressons notamment à la désambiguïsation d'entités spatiales dans le but d'effectuer des traitements plus avancés comme la géolocalisation ou l'inférence spatiale [Caron et al., 2012]. Cette étape est réalisée en associant un identifiant GeoNames [100] unique (une URI) à chaque lieu extrait. 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 utilisons, plus précisément, les trois propriétés suivantes pour l'agrégation des événements (voir la section 6.3.3) :

– 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.

La section 6.3.3 détaille comment nous utilisons ces relations topologiques pour l'agrégation des entités spatiales.

4.2.4 La dimension agentive

Comme dit précédemment, tous les participants d'un événement et leurs rôles respectifs sont représentés formellement par la dimension 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 $j \in \mathbb{N}$. 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 différents

100. http://www.geonames.org/

types de rôles possibles ne sont également pas restreints ici car cela est fortement corrélé à l'application et aux capacités des extracteurs. Notre méthode d'extraction (voir la section 5.4) n'ayant pas été conçue pour déterminer le rôle joué par un participant dans un événement, cet aspect ne sera pas traité ici.

4.3 WOOKIE : une ontologie dédiée au ROSO

Afin de définir précisément quelles sont les informations d'intérêt pour notre application, nous avons développé une ontologie de domaine nommée WOOKIE constituant la base de notre système de capitalisation des connaissances. Celle-ci a été créée de façon collaborative et a pour but de représenter les concepts de base nécessaires aux diverses applications du domaine. Cet aspect collaboratif (grâce notamment au logiciel WebProtégé) constitue un point important car il nous a permis de mettre en commun les visions de plusieurs membres de l'équipe IPCC afin de mieux cerner les besoins des opérationnels du renseignement. De plus, cette ontologie est implémentée au format OWL [101] selon les recommandations du W3C [102] pour la représentation des connaissances au sein du Web Sémantique. Notre objectif étant de développer une ontologie à taille raisonnable mais générique pour le renseignement militaire, nous avons tout d'abord mené quelques recherches pour faire le point sur les ontologies existantes. En effet, il nous est apparu intéressant de pouvoir reprendre tout ou partie d'une modélisation déjà disponible. WOOKIE a donc été élaborée suite à un état de l'art approfondi de la littérature du domaine et des ontologies existant à l'heure actuelle (dont une partie est présentée dans le chapitre 1).

Nous avons commencé par examiner les ontologies générales (dites "de haut niveau") les plus connues et utilisées telles que SUMO [103], PROTON [104], BFO [105], DOLCE [106] ou encore COSMO [107]. Ces ontologies sont modélisées à divers niveaux : elles définissent des concepts de haut niveau ou "meta-concepts" (pouvant servir de base à l'organisation d'encyclopédies par exemple) mais aussi des spécialisations des concepts *Lieu* et *Organisation* qui se sont avérées particulièrement intéressantes pour le développement de notre ontologie. Puis, d'autres modélisations plus spécifiques nous ont été utiles, telles que des modélisations spécifiques au domaine militaire (voir la section 1.3.3), des ontologies spécialisées dans la description des évènements (voir la section 1.3). Ces différentes observations ont montré qu'aucune des ontologies trouvées ne correspondaient parfaitement au modèle de connaissances voulu et qu'il n'était donc pas adéquat de reprendre une de ces représentations en l'état. Toutefois, nous nous sommes inspirés de ces modélisations tout au long de la construction de WOOKIE et avons veillé à maintenir des équivalences sémantiques avec les ontologies existantes. Le développement de notre ontologie de domaine a été guidé par les méthodologies de la littérature telles que [Noy and Mcguinness, 2001] [Mizoguchi, 2003a]. Celles-ci nous ont permis d'organiser notre travail en plusieurs étapes que nous détaillons ci-après.

Nous avons commencé par concevoir une taxonomie de concepts de haut niveau constituant la base de notre modélisation. Pour cela, nous avons accordé une attention particulière aux standards OTAN car ils définissent les objets d'intérêt et leurs propriétés pour le renseignement militaire. Ces standards demeurant trop techniques et détaillés, nous avons fait le choix de nous concentrer sur les catégories de

101. Ontology Web Language, http://www.w3.org/TR/owl-features/
102. World Wide Web Consortium, http://www.w3.org/
103. Suggested Upper Merged Ontology
104. PROTo ONtology
105. Basic Formal Ontology
106. Descriptive Ontology for Linguistic and Cognitive Engineering
107. Common Semantic MOdel

l'intelligence définies par le STANAG 2433, connues sous le nom de "pentagramme du renseignement" (voir la figure 4.1).

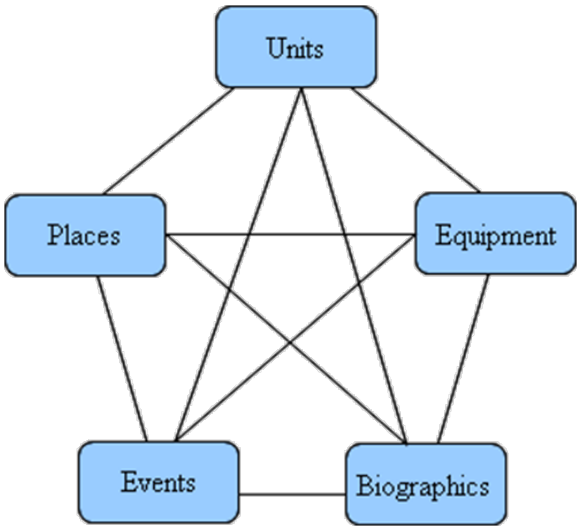


FIGURE 4.1 – Le pentagramme du renseignement

Ce pentagramme reprend les éléments centraux du domaine du renseignement militaire et les définit comme ci-dessous.

- Le concept *Units* y est défini comme tout type de rassemblement humain partageant un même objectif, pouvant être hiérarchiquement structuré et divisé en sous-groupes. Il s'agit à la fois des organisations militaires, civiles, criminelles, terroristes, religieuses, etc.
- Le concept *Equipment* désigne toute sorte de matériel destiné à équiper une personne, une organisation ou un lieu pour remplir son rôle. Il peut s'agir d'équipement militaire ou civil, terrestre, aérien, spatial ou sous-marin.
- Le concept *Places* regroupe les points ou espaces terrestres ou spatiaux, naturels ou construits par l'homme, pouvant être désignés par un ensemble de coordonnées géographiques.
- Le concept *Biographics* désigne les individus et décrit un certain nombre de propriétés associées telles que des éléments d'identification, des informations sur la vie sociale et privée, un ensemble de relations avec d'autres individus ou organisations, etc.
- Le concept *Events* décrit toute occurrence d'un élément considéré comme ayant de l'importance. Un événement peut être divisé en plusieurs sous-événements.

Toutefois, les standards OTAN ne détaillent pas les sous-classes du pentagramme et les diverses propriétés de classes évoquées doivent être triées et réorganisées. Nous avons donc développé notre ontologie de haut en bas (approche descendante ou "top-down"), en partant des concepts plus généraux vers les plus spécifiques. Nous avons effectué cette spécialisation en conservant les classes intéressantes des autres ontologies observées. Pour ce faire, nous nous sommes inspirés de la méthodologie de construction d'une ontologie proposée par [Noy and Mcguinness, 2001].

La création de sous-classes a été guidée par le contexte du renseignement militaire et ses éléments d'intérêt. La taxonomie complète des classes de WOOKIE est donnée en annexe A. Pour la classe *Equipment*, nous nous sommes limités à décrire les différents types de véhicules et d'armes. La classe *Person*

est liée par équivalence au concept *Person* dans l'ontologie FOAF [108] mais n'a pas nécessité plus de précision. En ce qui concerne le concept *Unit*, nous avons choisi de distinguer deux sous-classes *Group* et *Organisation* afin de différencier les groupements de personnes, point important dans le domaine visé. La classe *Place* a également été sous-typée en prenant en compte les besoins militaires, notamment par l'aspect stratégique des sous-classes du concept *Infrastructure*. Enfin, la modélisation de la classe *Event* s'est avérée une tâche essentielle compte tenu de l'importance de ces entités dans le renseignement et la veille militaire. Nous avons pour cela réservé plus de temps à la spécification de la classe *MilitaryEvent*, c'est-à-dire au choix et à l'organisation des différentes sous-classes en prenant en compte les observations préalables. La taxonomie des événements spécifiques au ROSO est présentée en annexe B.

Par la suite, parallèlement aux relations hiérarchiques, nous avons liés les concepts de l'ontologie entre eux par des relations sémantiques (*object properties*). Il s'agit de propriétés ayant pour co-domaine un concept de l'ontologie. Celles-ci ont également été choisies en fonction des besoins du renseignement militaire et en concertation avec les membres de notre équipe. Plusieurs liens sont modélisés entre les personnes tels que des liens familiaux, de connaissance, des liens hiérarchiques, etc. (*isFamilyOf*, *isSpouseOf*, *isFriendOf*, *isColleagueOf*, etc.). Nous avons également créé des relations entre personnes et organisations (*hasEmployee*), entre organisations et équipements (*producesEquipment*, *sellsEquipment*) ou encore entre lieux et personnes (*bornIn*, *diedIn*), etc. Enfin, la classe *Event* est en relation avec tous les autres éléments du pentagramme conformément à notre modèle d'événement (voir la section 4.2). Ainsi par exemple, un événement implique des participants de type *Person* ou *Unit* (*involves*) ainsi qu'un instrument appartenant à la classe *Equipment* (*hasEquipment*) et se déroule dans un lieu associé à la classe *Place* (*takesPlaceAt*). Un événement peut également être relié à d'autres événements par des relations d'antécédence, succession, cause, conséquence, etc. (*hasAssociatedEvent*, *causes*, *follows*, etc.).

Une quatrième étape a été d'attribuer à chaque classe un ensemble de propriétés permettant de les définir plus précisément en leur associant une valeur particulière. Cette valeur peut être une chaine de caractères, un nombre, une date, un booléen, etc. Comme nous l'avons déjà précisé plus haut, ces attributs sont héréditaires : ceux de la classe-mère sont automatiquement transmis aux classes-filles. Les 5 classes de plus haut niveau possèdent les attributs *picture*, pour leur associer une image, et *alias*, qui associé à la propriété *rdfs :label* permet d'indiquer leur(s) nom(s) alternatif(s). La classe *Person* possède un certain nombre d'attributs tels que la nationalité, la profession, l'âge, l'adresse postale, l'adresse électronique, les dates de naissance et de décès, etc. La classe *Unit* est également caractérisée par des adresses postale et électronique ainsi que des coordonnées téléphoniques. La classe Place n'a pas nécessité d'attributs. Le concept *Equipment* possède lui des attributs essentiellement liés à des caractéristiques techniques telles que la couleur, les dimensions, la vitesse ou encore à des informations d'identification comme la marque, le modèle, la plaque d'immatriculation, l'année de production, etc. Enfin, pour la classe *Event*, nous avons précisé les dates de début et fin, la durée ainsi que le nombre de victimes et de décès engendrés. La totalité des attributs de concepts est donnée en annexe D).

Pour terminer, nous avons précisé les différents propriétés et attributs définis en spécifiant certaines contraintes et axiomes dans l'ontologie WOOKIE. Comme permis par la spécification OWL, des liens entre propriétés de type *owl :inverseOf* ont été implémentés pour indiquer que telle propriété porte un sens contraire à telle autre propriété. Par exemple, la relation *isEmployeeOf* (qui lie une personne à l'organisation à laquelle elle appartient) est l'inverse de *hasEmployee*. Par ailleurs, nous avons utilisé, le cas échéant, les restrictions *symmetric/assymetric*, *irreflexive* et *transitive* des modèles OWL et OWL2. La propriété *hasFriend* est, par exemple, symétrique et non-réflexive. La relation *isSuperiorTo* entre

108. Friend Of A Friend, http://www.foaf-project.org/

deux membres d'une organisation est quant à elle transitive. Enfin, comme mentionné en section 1.3.1.2, WOOKIE intègre des liens sémantiques avec d'autres ontologies sous forme d'axiomes tels que des équivalences de classes (*owl :equivalentClass*) ou des relations de subsomption (*rdfs :subClassOf*). Ainsi, le concept *Person* de WOOKIE est équivalent au concept *Person* de l'ontologie FOAF et la classe *Place* de WOOKIE est un sous-type de *Feature* dans l'ontologie GeoNames. Par ailleurs, le concept d'événement dans notre ontologie équivaut sémantiquement au concept *Event* de l'ontologie LODE, par exemple.

4.4 Conclusions

L'ontologie que nous venons de décrire constitue le modèle de connaissances qui servira de guide à notre approche d'extraction et de capitalisation des connaissances. Les principaux atouts de cette modélisation sont les suivants : tout d'abord, elle se fonde sur des modèles reconnus (ACE et DUL pour les événements et le modèle TUS et les relations topologiques RCC-8 pour la représentation spatio-temporelle). De plus, notre ontologie a été conçue en accord avec les besoins du ROSO (taxonomie de classes et propriétés) et intègre de nombreux liens sémantiques vers d'autres ontologies afin de maintenir une interopérabilité au sein du Web sémantique. Celle-ci présente toutefois quelques limites et nous envisageons des perspectives d'amélioration telles que l'intégration d'une cinquième dimension que nous appellerons "contextuelle" afin de représenter des éléments du contexte linguistique et extra-linguistique (indices de modalité, confiance, temporalité, propagation spatiale, etc.). Par ailleurs, nous souhaitons approfondir la représentation des rôles au sein de la dimension agentive en étudiant, par exemple, le modèle SEM (voir le chapitre 1.3.1.2).

Chapitre 5

Extraction automatique des événements

Sommaire

5.1 Introduction

Une fois notre modèle de connaissances établi, nous proposons de concevoir une approche permettant de reconnaitre automatiquement dans un ensemble de textes les différentes informations d'intérêt pour peupler la base de connaissances (créer des instances des différentes classes de l'ontologie WOOKIE et les liens existants entre ces instances). Ce chapitre présente notre seconde contribution : un système d'extraction automatique d'événements pour la veille en sources ouvertes. Celui-ci ayant été essentiellement élaboré grâce à la plateforme GATE, nos critères de choix ainsi qu'une présentation générale de cet outil sont exposés dans une première partie. Nous décrivons par la suite le développement d'un extracteur d'entités nommées nécessaire à la reconnaissance des événements tels que définis dans le chapitre 4.2. Nous terminons par une présentation de notre système de reconnaissance des événements fondé sur deux méthodes : une approche à base de règles linguistiques, d'une part, et une méthode par apprentissage de motifs fréquents, d'autre part. Les paramètres choisis pour notre application ainsi que les performances de notre système d'EI seront présentés en section 7.2.

5.2 La plateforme GATE

GATE est une plateforme *open-source* implémentée en Java dédiée à l'ingénierie textuelle au sens large [Cunningham et al., 2002]. Créée il y a une vingtaine d'années par les chercheurs de l'université de Sheffield (Royaume-Uni), GATE est largement utilisée par les experts en TAL et dispose d'une grande communauté d'utilisateurs. Cela lui permet de proposer un ensemble de solutions d'aide et de support (forum, liste de diffusion, foire aux questions, wiki, tutoriels, etc.). Ce point a constitué un critère important afin de choisir notre environnement de développement parmi les différentes plateformes et outils présentés en section 2.3. Par ailleurs, GATE propose une chaine d'extraction d'entités nommées pour l'anglais nommée ANNIE composée de différents modules *open source*. Cette chaine ayant déjà été utilisée au sein de la plateforme WebLab, elle a constitué une première base pour l'élaboration de notre propre système d'extraction d'information.

Fonctionnement général

L'environnement GATE repose sur le principe de chaines de traitement composées de différents modules (dits "Processing Resources" PR) appliqués successivement sur un ou plusieurs textes (dits "Language Resources" LR). Les LR peuvent être des textes seuls, fournit dans l'interface par copier-coller ou URL, ou des corpus de textes créés manuellement ou importés d'un dossier existant. Produit d'une communauté d'utilisateurs croissante, l'ensemble des PR disponibles est conséquent : ceux-ci sont organisés au sein de *plugins* thématiques et permettent des traitements variés pour une quinzaine de langues au total. L'utilisateur peut ainsi sélectionner un ensemble de briques logicielles pertinentes pour sa tâche, les paramétrer et les organiser à sa convenance pour construire une chaine de traitement de texte adaptée (voir l'annexe E pour un exemple de chaine de traitement).

La majorité des modules de traitement fournit une analyse des textes à traiter par un système d'annotations exprimées au format XML et selon le modèle de la plateforme. Ce système permet à chaque brique de traitement d'exploiter les annotations fournies par les modules précédents. Cela consiste géné-

ralement à associer un ensemble d'attributs à une zone de texte. Ces attributs se présentent sous la forme *propriété = "valeur"*. Voici un exemple d'annotation créée par un composant de segmentation en mots :

```
Token {category=NNP, kind=word, length=5, orth=upperInitial,
string=Obama}
```

Il s'agit ici d'une annotation de type *Token* à laquelle sont associés plusieurs attributs comme sa catégorie grammaticale (*NNP*), son type (*word*), sa longueur en nombre de caractères (*5*), sa casse (*upperInitial*) et la chaîne de caractères correspondante (*Obama*).

Le formalisme JAPE

Parallèlement aux différents modules d'analyse, la plateforme GATE propose un formalisme d'expression de grammaires contextuelles nommé JAPE (Java Annotation Patterns Engine). Ce formalisme est employé pour la définition de règles au sein des modules de type transducteurs à états finis. Ce système s'avère très utile en extraction d'information car il permet de définir les contextes d'apparition des éléments à extraire pour ensuite les repérer et les annoter dans un ensemble de textes. Le principe est de combiner différentes annotations fournies par les modules précédant le transducteur (*tokens*, syntagmes, relations syntaxiques, etc.) pour en créer de nouvelles plus complexes (entités nommées, relations, évènements, etc.) : cela revient à l'écriture de règles de production et donc à l'élaboration d'une grammaire régulière. Une grammaire dans GATE se décompose en plusieurs phases exécutées consécutivement et formant une cascade d'automates à états finis. Chaque phase correspond à un fichier *.jape* et peut être constituée d'une ou plusieurs règle(s) écrite(s) selon le formalisme JAPE. Classiquement, ces règles sont divisées en deux blocs : une partie gauche (*Left Hand Side* ou *LHS*) définissant un contexte d'annotations à repérer et une partie droite (*Right Hand Side* ou *RHS*) contenant les opérations à effectuer sur le corpus à traiter lorsque le contexte *LHS* y a été repéré. Le lien entre ces deux parties se fait en attribuant des étiquettes à tout ou partie du contexte défini en *LHS* afin de cibler les annotations apposées en *RHS*. Pour plus de clarté, prenons l'exemple d'une règle simple :

```
1 Rule: OrgAcronym
2 (
3         {Organisation}
4         {Token.string == "("}
5         ({Token.orth == "allCaps"}):org
6         {Token.string == ")"}
7 )
8 -->
9 :org.Organisation = {rule="OrgAcronymRule", kind="Acronym"}
```

FIGURE 5.1 – Exemple de règle d'extraction exprimée dans le formalisme JAPE

L'objectif de celle-ci est d'annoter avec le type *Organisation* tous les acronymes entre parenthèses positionnés après une première annotation de ce même type. La première ligne de cette règle indique le nom qui lui a été donné par son auteur. Les lignes 2 à 7 définissent le motif à repérer dans le texte : les types des annotations sont encadrés par des accolades (e.g. *{Organisation}*), l'accès aux attributs

des annotations se fait sous la forme *Annotation.attribut* (*Token.string*, par exemple, permet d'obtenir la valeur de la propriété *string* associée aux annotations de type *Token* fournies par un module précédent), *(...) :org* permet d'étiqueter la partie visée du motif pour y référer en partie *RHS* de la règle. Puis, la flèche (ligne 8) sert de séparateur entre les parties *LHS* et *RHS*. La ligne 9 permet d'attribuer une annotation de type *Organisation* au segment étiqueté *org* en partie gauche. Remarquons également en partie droite l'ajout des propriétés *rule* et *kind* à l'annotation produite afin d'indiquer quelle règle en est à l'origine et qu'il s'agit d'un acronyme d'organisation. Ces attributs peuvent être librement définis par le développeur de grammaires.

Précisons également qu'un système de macros permet de nommer une séquence d'annotations afin de la réutiliser de façon raccourcie dans les règles définies. Enfin, il est possible de gérer l'ordre d'exécution d'un ensemble de règles en choisissant un des différents modes de contrôle proposés par le formalisme JAPE (*all*, *once*, *appelt*, etc.).

Quelques modules utilisés dans nos travaux

Dans cette section, nous présentons différents modules qui nous ont été utiles pour mettre en oeuvre notre système d'extraction d'information. Tout d'abord, nous devons parler d'ANNIE (A Nearly-New Information Extraction system), une chaine complète dédiée à la reconnaissance d'entités nommées pour l'anglais. Fournie conjointement à la plateforme, cette chaine comprend différents modules d'analyse :

1. *Document Reset* : supprime toutes les annotations apposées précédemment sur le document,
2. *English Tokenizer* : découpe le texte en mots (*tokens*),
3. *Gazetteer* : repère les éléments contenus dans une liste (*gazetteer*) et les annote en tant que *Lookup*,
4. *Sentence Splitter* : découpe le texte en phrases,
5. *POS Tagger* : ajoute à l'annotation *Token* (mise par le *tokenizer*) une propriété *category* indiquant la catégorie morpho-syntaxique du mot en question. Il s'agit, ici, d'une version dérivée de l'étiqueteur Brill [Brill, 1992],
6. *NE Transducer* : un transducteur JAPE définissant un ensemble de règles afin de repérer des entités nommées (*Person*, *Organization*, *Location*, *Date*, *URL*, *Phone*, *Mail*, *Address*, etc.),
7. *OrthoMatcher* : annote les relations de co-référence entre les entités nommées repérées précédemment.

ANNIE étant spécialisée pour le traitement de textes anglais, d'autres modules se sont avérés nécessaires pour l'extraction d'information en français. Nous avons notamment utilisé les modules de découpage en mots et d'étiquetage morpho-syntaxique adaptés pour la langue française (*French tokenizer* et *TreeTagger*). De plus, la phase d'extraction d'événements a nécessité l'utilisation de l'analyseur syntaxique *Stanford parser* fournissant une analyse syntaxique en dépendance des phrases à traiter. Par ailleurs, nos travaux nécessitant un découpage des textes en groupes syntaxiques (nominaux et verbaux), les modules *NP Chunker* et *VP Chunker* ont répondu à ce besoin. Enfin, divers modules d'analyse linguistique nous ont été utiles tels qu'un analyseur morphologique ou encore un repérage lexical paramétrable (dit *Flexible Gazetteer*).

5.3 Extraction d'entités nommées

La phase d'extraction d'entités nommées consiste à mettre en place un système de détection et de typage des entités d'intérêt pour l'anglais et le français. En effet, pour reconnaitre des événements tels qu'ils sont définis dans WOOKIE, notre premier objectif est de repérer les entités de type *Person*, *Unit*, *Date* et *Place*. Pour ce faire, comme nous l'avons montré dans l'état de l'art correspondant (voir la section 2.2.1), plusieurs types de méthodes ont été explorées : des approches symboliques, statistiques ou encore hybrides. Celles-ci ayant montré des performances comparables pour le problème de la REN, notre choix a été guidé par le contexte applicatif de nos travaux. En effet, il apparait important dans le contexte du ROSO d'éviter l'extraction d'informations erronées (ce que l'on nomme plus communément le bruit). Ce critère nous a donc orienté vers le choix d'une approche à base de règles, pour laquelle la littérature a montré une plus grande précision et qui s'avère également plus facilement adaptable à un domaine d'application donné.

Notre système de REN a été élaboré selon un processus ascendant et itératif : nous avons collecté un ensemble de dépêches de presse en anglais et français et repéré manuellement des exemples d'entités nommées à extraire. Partant de ces exemples, nous avons construit (par généralisations successives des contextes d'apparition de ces entités) une première version du système, que nous avons appliqué sur ce même jeu de textes pour vérifier la qualité des règles construites (en termes de précision et rappel). Le système a ensuite été modifié pour atteindre la qualité voulue, puis un nouveau corpus de textes a été constitué pour découvrir de nouvelles formes d'entités et ainsi de suite. De par la proximité des langues anglaise et française, cette méthode a été appliquée pour les deux langues et les systèmes d'extraction développés suivent donc les mêmes principes et présentent une structure commune. Nous présentons ci-dessous les différentes étapes d'extraction ainsi que leur implémentation pour l'anglais et le français grâce à la plateforme GATE.

5.3.1 Composition de la chaine d'extraction

La chaine d'extraction d'entités nommées est composée de différents modules d'analyse listés et décrits ci-après. L'ordre d'exécution de ces modules a son importance car chaque traitement exploite les annotations créées précédemment. Pour cela, nous nous sommes inspirés de la chaine d'extraction ANNIE présentée ci-dessus tout en l'adaptant à notre problématique et à notre représentation des connaissances. La composition de modules obtenue reste la même pour l'anglais et le français bien que certains des modules soient spécifiques à la langue traitée. Notons également que les quatre premières étapes sont réalisées par des briques (reprises en l'état) proposées dans GATE tandis que les deux derniers modules ont été au centre de nos travaux et ont nécessité notre expertise.

1. *Réinitialisation du document*
   Ce premier module permet de nettoyer le document traité de toute annotation existante afin de le préparer à l'exécution d'une nouvelle chaine d'extraction. Il est indépendant de la langue.
2. *Découpage en mots*
   Ce second module découpe le texte en mots ou plus précisément en *tokens*. Ce traitement dépend de la langue traitée : nous avons donc utilisé les *tokenizers* spécifiques à l'anglais et au français fournis dans GATE.

3. *Découpage en phrases*

   Cette étape permet d'obtenir une annotation par phrase du texte. L'anglais et le français étant des langues proches syntaxiquement, le même module a été utilisé pour ces deux langues.

4. *Étiquetage grammatical*

   Également nommée "étiquetage morpho-syntaxique" ou *Part-Of-Speech (POS) tagging* en anglais, cette phase donne pour chaque *token* repéré un ensemble d'informations grammaticales telles que la catégorie grammaticale (nom, verbe, adjectif, etc.), le genre, le nombre, etc. Ce module est dépendant de la langue considérée : l'analyseur pour l'anglais est celui présent dans la chaine ANNIE tandis que pour le français nous avons choisi l'outil TreeTagger.

5. *Repérage lexical*

   Cette étape consiste à repérer dans les textes à traiter un ensemble de mots clés préalablement définis dans des listes nommées *gazetteers*. Ces mots clés sont généralement des termes communs dont l'occurrence peut indiquer la présence d'une entité nommée. Les *gazetteers* sont de nature variée : listes de prénoms pour la reconnaissance des noms de personnes, jours de la semaine et noms des mois pour la détection des dates, etc. A chaque liste est associée une étiquette sémantique qui constituera le type de l'annotation apposée. Même s'il s'agit généralement d'une simple projection de lexique, ce module s'avère très utile pour l'extraction des entités nommées en tant que telle effectuée par le module suivant. En effet, les annotations fournies par ce module entrent dans la définition des différents contextes d'apparition des entités nommées (partie gauche des règles). Un exemple de *gazetteer* pour la détection des personnes en français est présentée en annexe F.

6. *Règles d'extraction*

   Ce dernier module d'annotation constitue le cœur du système de REN et contient les règles linguistiques élaborées manuellement, suivant le formalisme JAPE, pour la détection des entités de type *Personne*, *Organisation*, *Lieu* et *Date*.

Nous avons fait le choix, dans le cadre de nos recherches, de ne pas conserver le module de résolution de co-référence *Orthomatcher* au sein de notre système d'extraction. En effet, nous avons constaté que la précision de ce module n'étant pas suffisamment bonne pour notre cas d'application et que cela détériorait la qualité de l'extraction d'événements (fortement dépendantes de l'extraction d'entités nommées).

5.3.2 Développement du module de règles linguistiques

Le module d'extraction constitue le cœur de notre système. L'ensemble des règles linguistiques développées sont contextuelles et partagent la forme suivante :

$$regle : contexte \rightarrow action$$

Elles sont implémentées selon le formalisme JAPE décrit en section 5.2.

Conformément à notre cas d'application, nous avons veillé, lors de leur développement, à privilégier la précision au rappel, c'est-à-dire l'extraction d'informations pertinentes pour mieux répondre à l'attente des opérationnels du domaine. Concrètement, cela s'est traduit par la construction de règles linguistiques dont les résultats sont plus sûrs et la mise à l'écart de règles pouvant entrainer de fausses annotations.

Celles-ci sont organisées en plusieurs ensembles, chacun étant spécifique au type d'entité ciblé. Ces ensembles sont implémentés sous la forme d'automates à états finis et l'exécution des règles au sein d'un même ensemble est régie par un système de priorité. Celui-ci permet de déterminer, lorsqu'il y a conflit entre plusieurs règles, quelle est celle qui doit être privilégiée pour annoter une portion de texte. Les différents ensembles de règles sont eux exécutés successivement selon un ordre fixé à l'avance au sein du système, constituant ce que l'on appelle des phases d'extraction. Pour déterminer le meilleur agencement de ces phases au sein du module de règles, nous nous sommes référés aux travaux de [Mikheev, 1999] et plus particulièrement à la gestion d'éventuelles ambiguïtés entre entités. Le principe suggéré est le suivant : afin d'éviter toute ambiguïté, il est conseillé d'exécuter en premier les règles les plus sûres (basées essentiellement sur le contexte linguistique), puis de typer les entités encore inconnues grâce aux *gazetteers* du module précédent et de lever les ambiguïtés restantes dans une phase finale. Nos propres observations nous ont amenés à choisir, en outre, un ordre de détection entre les 4 entités-cibles : nous typons, tout d'abord, les dates qui ne sont généralement pas confondues avec les autres entités ; puis, vient une phase de détection des organisations, suivie par les entités de type *Personne* et, enfin, les noms de lieux. En effet, nous avons observé que les noms d'organisations peuvent inclure des noms de personnes ou de lieux et doivent donc être repérés en priorité afin d'écarter l'ambiguïté. Par ailleurs, certains prénoms présentant une homonymie avec des noms de lieux, les entités de type *Personne* doivent être extraites avant celles de type *Lieu*. Ces premières présentent des formes moins variables et sont donc plus facilement repérables (grâce notamment aux listes de prénoms et de titres personnels).

Pour l'extraction des EN en anglais, nous sommes partis de l'existant (la chaine ANNIE) en modifiant l'ordre des phases selon le principe explicité ci-dessus et sélectionnant/améliorant les règles les plus pertinentes pour notre application. Dans le cas du français, nous avons construit le système complet en s'appuyant sur la même méthodologie. Nous présentons, pour conclure cette section, quelques exemples de règles et *gazetteers* pour chaque type d'extraction.

Extraction des dates

La référence au temps peut s'exprimer de façon diverse et les expressions temporelles revêtent des formes textuelles variées : littérales ("en janvier"), numériques (10/01/2013) ou mixtes ("le 9 janvier"). La littérature distingue les expressions dites "absolues" de celles dites "relatives" : les premières permettent à elles seules de se repérer sur un axe temporel (par exemple, "le 9 janvier 2002" ou "l'année 2010") alors que les secondes ("hier matin", "le 5 mars", etc.) nécessitent pour cela des informations complémentaires provenant du co-texte ou du contexte extra-linguistique. On constate aussi des différences dans l'expression des dates d'une langue à une autre : dans notre cas, les anglophones n'expriment pas les dates comme les francophones. Toutes ces variations rendent la détection automatique des dates complexe et le développement des règles doit être réalisé en tenant compte des spécificités de l'anglais et du français.

Voici pour exemple deux règles JAPE extraites de notre système : la première 5.2 permet d'extraire des dates en français, la seconde 5.3 est l'équivalent pour l'anglais (les dates en gris sont des exemples pouvant être détectés par la règle). Comme mentionné plus haut, il est fait usage de macros (termes en majuscules) pour faciliter la lecture et la maintenance du jeu de règles. De plus, un ensemble de propriétés est adjoint à l'annotation pour faciliter la normalisation des dates (voir la section 6.2).

```
// lundi 27 avril 2009
// 27 avril 2009
// 1er avril 2006
Rule: DateComplete
(
    (NOM_JOUR)?
    (NB_JOUR):jour
    ({Token.string == "-"})?
    (NOM_MOIS):mois
    (ANNEE):annee
):date
-->
:date.Date = {rule = "DateComplete", startYear = :annee.Token.string,
startMonth = :mois.Lookup.value, startDay = :jour.Lookup.value}
```

FIGURE 5.2 – Règle d'extraction de dates en français

```
// Wed 10 July, 2000
// Sun, 21 May 2000
// 10th of July, 2000
Rule:      DateNameComplete
Priority: 100
(
    (DAY_NAME (COMMA)?)?
    (ORDINAL | DAY_NUM):day
    (MONTH_NAME):month
    (COMMA)?
    (YEAR):year
):date
-->
:date.Date = {rule = "DateNameComplete", startYear =
:year.Token.string, startMonth = :month.Lookup.value, startDay =
:day.Token.string}
```

FIGURE 5.3 – Règle d'extraction de dates en anglais

Extraction des organisations

Le système développé permet d'extraire des noms d'organisations divers tels que : *NATO, Communist Party of India-Maoist, Ouattara Party, Nations Unies, AFP, Radio Azzatyk, armée de l'Air, etc.* Nous utilisons pour cela des gazetteers de mots couramment associés à des noms d'organisations, qu'ils fassent partie de l'entité nommée ou non (la figure 5.4 est un extrait d'un de ces *gazetteers* pour l'anglais).

La règle ci-dessous 5.5 utilise un de ces gazetteers afin de reconnaitre des noms d'organisations en anglais précédés d'un titre ou d'une fonction de personne tels que *the director of the FBI* ou *the interim vice president of Comcast Business*.

Extraction des personnes

Concernant la reconnaissance des noms de personne, nous avons assez classiquement utilisé des *gazetteers* de prénoms. Toutefois, face à l'immense variété des prénoms (même en domaine restreint) et

```
Honorary Society
Horse
Horse Cavalry
Hospital
Host
Hotel
Hotels
House
Household
Housing Industry
Hunt
Hunt Club
INC
INCORPORATED
INDUSTRIES
INSTITUT
INSTITUTE
INSTITUTES
Inc
Incorporated
Index Fund
Industrial Bank
Industrial Loan Company
Industrial Union
Industries
Industry
Infant School
Infantry
Institut
Institute
Institutes
Institution
```

FIGURE 5.4 – Extrait du gazetteer *org_key.lst*

```
Rule: OrgTitle
Priority: 60
(
   {Lookup.majorType == "jobtitle"}
   {Token.string == "of"}
   ({Token.string == "the"})?
   (
      (UPPER (POSS)?)[1,4]
      ORG_KEY
   ):org
)
-->
:org.Organisation = {rule = "OrgTitle"}
```

FIGURE 5.5 – Règle d'extraction d'organisations en anglais

les ambiguïtés possibles, il est nécessaire d'exploiter d'autres indices contextuels tels que les titres ou fonctions personnelles. Ces derniers peuvent être placés en début ou en fin de l'entité nommée, la figure 5.6 donne quelques exemples d'indices antéposés en anglais.

Enfin, la règle suivante 5.7, par exemple, exploite les annotations posées par différents *gazetteers* (fonctions et nationalités ici) pour l'extraction des noms de personne.

```
Maj-Gen.
Maj.
Maj. Gen
Major
Major General
Major-General
Manager
Managing Director
Maréchal
Marquis
Marshal
Marshal Of The R.A.F.
Marshal Of The RAF
Marshal of the R.A.F.
Marshal of the RAF
Master Sergeant
Master Serjeant
Master Sgt.
Mayor
Md
Messr
Messr.
Messrs
Messrs.
Midshipman
Midshipwoman
Minister
Miss
Mlle
Mme
Mme.
Monsieur
Mr
Mr.
Mrs
Mrs.
Ms
Ms.
Nurse
Officer
```

FIGURE 5.6 – Extrait du gazetteer *person_pre.lst*

```
Rule: PersFonction
(
   FONCTION
   (NATIONALITE)?
   (((NP)[1,3]):last):person
)
-->
:person.Person = {rule = "PersFonction"}
```

FIGURE 5.7 – Règle d'extraction de personnes en français

Extraction des lieux

L'extraction automatique des noms de lieux est, avec celle des organisations, l'une des plus complexe à mettre en œuvre de par la grande diversité formelle et référentielle de ces entités. En effet, les dépêches de presse relatant généralement des faits de façon précise, nous pouvons y trouver des noms de lieux granularité variable (noms de pays, villes, villages, régions, quartiers, rues, etc.) et exprimés parfois de façon relative ou imprécise (par exemple, *eastern Iraq, Asie centrale, au nord de l'Afghanistan*). De même que

pour la détection des organisations, de nombreuses ressources géographiques sont disponibles librement (*gazetteers* de toponymes, bases de données géographiques, etc.) et nous avons pu constituer quelques listes de base pour la détection des lieux communément mentionnés (liste des pays, des continents, des capitales, etc.). Il est nécessaire également de fonder nos règles d'extraction sur des indices contextuels tels que des noms communs déclencheurs d'entités géographiques (voir la figure 5.8 pour des exemples en français).

```
école
église
état
île
Avenue
Basilique
Bibliothèque
Bois
Boulevard
Cathédrale
Collège
Ecole
Eglise
Gare
Hôtel
Hôtel de ville
Hippodrome
Institut
Jardin
Jardins
Lycée
Maison
Mosquée
Opéra
Palais
Parc
Place
Pont
Quartier
Rue
Théâtre
aéroport
arrondissement
```

FIGURE 5.8 – Extrait du gazetteer *loc_key.lst*

Pour finir, la figure 5.9 est un exemple de règle JAPE exploitant ces indices de contexte :

```
Rule: LocKeyIncl
Priority: 50
(
    {Lookup.minorType == "loc_key_incl"}
    (DE)?
    (ARTICLE)?
    (UPPER)[1,3](ADJLOC)?
):loc
-->
:loc.Location = {rule = "LocKeyIncl"}
```

FIGURE 5.9 – Règle d'extraction de lieux en français

5.4 Extraction d'événements

Nous détaillons, dans cette section, le cœur de notre première contribution à savoir la conception et l'implémentation d'un système d'extraction automatique d'événements tels que nous les avons définis au chapitre 4. L'objectif est donc d'extraire un ensemble d'événements associés aux participants et circonstants suivants : la date de l'événement, son lieu d'occurrence et les entités de type *Personne* et *Organisation* impliquées. Celui-ci est constitué de deux extracteurs suivant deux approches distinctes : une méthode symbolique à base de règles linguistiques élaborées manuellement et une approche par apprentissage de motifs séquentiels fréquents. En effet, l'état de l'art réalisé (voir le chapitre 2.2.3) n'ayant pas révélé d'approche nettement supérieure aux autres, la combinaison de plusieurs méthodes parait être la solution la plus pertinente. Élaborer un système composite permet de tirer le meilleur parti des approches actuelles en exploitant la complémentarité des différents types d'approche (statistique, symbolique, etc.).

Nous nous sommes tournés, dans le cadre de cette thèse, vers la combinaison d'un système à base de règles et d'un apprentissage symbolique pour plusieurs raisons. Tout d'abord, de par les besoins du ROSO, il est préférable de privilégier l'extraction d'informations fiables et précises et les approches symboliques répondent bien à ce besoin. Par ailleurs, afin d'améliorer les performances de ces techniques, il est apparu intéressant de les combiner avec un système d'apprentissage, dont le rappel est généralement meilleur. Nous nous sommes, dans un premier temps, orientés vers un système statistique tels que les CRFs, cependant après plusieurs recherches, nous n'avions pas à disposition un corpus d'apprentissage adapté (annoté en événements du domaine militaire/sécurité). Cela nous a donc mené vers le choix d'une méthode d'apprentissage faiblement supervisée telle que l'extraction de motifs fréquents qui ne nécessite pas de données annotées avec les entités-cibles.

Nous présentons ci-dessous les principes théoriques de chaque approche choisie ainsi que la façon dont nous les avons mises en œuvre pour le traitement de dépêches de presse en langue anglaise. La méthode élaborée se veut indépendante de la langue des textes considérés et a été illustrée, dans le cadre de cette thèse, pour des textes en anglais uniquement en raison d'une plus grande disponibilité pour cette langue des outils et données nécessaires.

5.4.1 Approche symbolique

La première méthode employée pour la détection d'événements est fondée sur la définition de règles linguistiques contextuelles (du même type que celles présentées en section 5.3) couplée avec une analyse syntaxique des textes. Celle-ci a été implémentée grâce à la plateforme GATE sous la forme d'une chaîne de traitement composée de différents modules d'analyse linguistique (*tokenisation*, découpage en phrases, repérage lexical, étiquetage grammatical, analyse syntaxique, etc.) et se déroule en plusieurs étapes détaillée ci-après.

Repérage des déclencheurs d'événements

La première étape consiste à repérer dans les textes à traiter les termes qui réfèrent potentiellement aux événements ciblés, que nous appellerons des "déclencheurs d'événements" (correspondant aux ancres du modèle ACE). Tout d'abord, nous considérons comme possibles déclencheurs d'événements

les verbes et les noms, éléments porteurs de sens. Le repérage de ces déclencheurs se fait par l'utilisation de *gazetteers* contenant, d'une part, des lemmes verbaux pour les déclencheurs de type *verbe* et, d'autre part, des lemmes nominaux pour les déclencheurs de type *nom*. Nous avons chois de constituer des listes de lemmes afin d'obtenir des listes plus courtes et d'étendre le repérage des déclencheurs à toutes les formes fléchies (grâce au module GATE de type *flexible gazetteer*). Ces déclencheurs (139 lemmes actuellement) ont été 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. Après une phase de découpage en mots, un analyseur morphologique attribue à chaque *token* son lemme. Nous comparons ensuite chaque lemme aux listes de déclencheurs et, s'ils correspondent, le mot lemmatisé est annoté comme étant un déclencheur d'événement. De plus, on lui associe la classe d'événement qu'il représente.

La figure 5.4.1 présente la liste des lemmes verbaux utilisés comme déclencheurs des événements de type *BombingEvent* (les informations indiquées après le symbole § seront clarifiées par la suite).

```
explode§struct=4
bomb§struct=1
detonate§struct=2
rock§struct=4
plant a bomb§struct=1
place a mine§struct=1
hurl§struct=1
dynamite§struct=1
```

FIGURE 5.10 – Gazetteer *bombings.lst*

Analyse des dépendances syntaxiques

Une fois les déclencheurs d'événement repérés, il nous faut leur associer les différentes entités impliquées pour former l'événement dans son ensemble. Pour cela, nous effectuons, dans un premier temps, une extraction automatique d'entités nommées grâce au système présenté en section 5.3 ainsi qu'une analyse en constituants syntaxiques (syntagmes nominaux *NP*, verbaux *VP*, prépositionnels *SP* ou adjectivaux *SA*).

Nous devons ensuite repérer les différentes relations entre le déclencheur et les entités de la phrase. Nous employons, dans ce but, un analyseur syntaxique donnant les dépendances entre les différents éléments phrastiques. En effet, une analyse syntaxique permet d'obtenir une meilleure précision par rapport à l'utilisation d'une analyse dite "par fenêtre de mots" associant un simple découpage en syntagmes (*chunking*) et des règles contextuelles. Après avoir examiné les différentes solutions proposées dans la plateforme GATE, nous avons opté pour l'utilisation du *Stanford parser* [De Marneffe and Manning, 2008]. Ici, nous faisons le choix de n'utiliser que les dépendances principales à savoir les relations *sujet*, *objet*, *préposition* et *modifieur de nom*. Les dépendances extraites par le *Stanford parser* se présentent sous la forme de liens entre les éléments centraux du syntagme-recteur et du syntagme-dépendant, plus communément appelés "têtes de syntagme" (voir la figure 5.11).

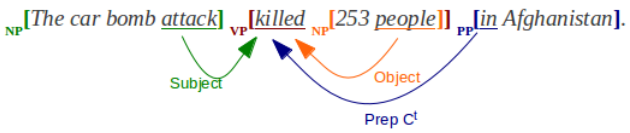


FIGURE 5.11 – Exemple d'analyse syntaxique en dépendance

	Voix active	Voix passive
Classe 1	sujet = agent, objet = patient	sujet = patient, ct. prep. "by" [109] = agent
Classe 2	sujet = agent, objet = instrument	sujet = instrument, ct. prep. "by" = agent
Classe 3	sujet = instrument, objet = patient	sujet = patient, ct. prep. "by" = instrument
Classe 4	sujet = agent, objet = locatif	sujet = locatif, ct. prep. "by" = agent
Classe 5	sujet = patient	∅

TABLE 5.1 – Classes argumentales pour l'attribution des rôles sémantiques

Attribution des rôles sémantiques

La dernière étape consiste à attribuer un rôle sémantique (agent, patient, instrument, etc.) aux différents participants de l'événement. Cela est rendu possible par une étude de la structure argumentale du verbe ou du nom déclencheur : à savoir déterminer sa valence (nombre d'arguments) et les rôles sémantiques de ses différents actants. Si nous prenons l'exemple des verbes anglais *kill* et *die*, nous remarquons qu'ils ont des valences différentes (2 et 1 respectivement) et que leurs sujets n'ont pas le même rôle sémantique : le premier sera *agent* et le second *patient*. L'attribution de ces rôles sémantiques nécessite d'étudier, en amont, la construction des lemmes présents dans nos *gazetteers* [François et al., 2007]. Pour cela, nous avons choisi de constituer 5 classes argumentales, chacune d'elles correspondant à un type de construction verbale ou nominale (voir le tableau 5.1).

L'attribution des rôles sémantiques est réalisée par un ensemble de règles linguistiques (implémenté sous la forme d'un transducteur JAPE) exploitant les informations fournies par les phases précédentes.

La figure 5.12 résume les différentes étapes de notre approche.

FIGURE 5.12 – Extraction des événements : différentes étapes

L'ensemble de ces traitements ont été implémentés grâce à des modules fournis dans GATE (voir la figure 5.13) et permettent d'obtenir une annotation positionnée sur l'ancre d'événement indiquant le type

de l'événement (sa classe dans l'ontologie de domaine) ainsi que les différentes entités impliquées (date, lieu et participants). Un exemple d'annotation obtenu est fourni par la figure 5.14.

!	Name	Type
	Document Reset	Document Reset PR
	Tokeniser_ENG	ANNIE English Tokeniser
	GazetteerNE_ENG	ANNIE Gazetteer
	Sentence Splitter	ANNIE Sentence Splitter
	POS Tagger_ENG	ANNIE POS Tagger
	Morphological analyser_ENG	GATE Morphological analyser
	FlexGazetteerEventNouns_ENG	Flexible Gazetteer
	FlexGazetteerEventVerbs_ENG	Flexible Gazetteer
	NP Chunker_ENG	Noun Phrase Chunker
	VP_Chunker_ENG	ANNIE VP Chunker
	StanfordParser_ENG	StanfordParser
	TransducerNE_ENG	ANNIE NE Transducer
	TransducerEvent_ENG	JAPE Transducer

FIGURE 5.13 – Extraction des événements : chaine de traitement GATE pour l'anglais

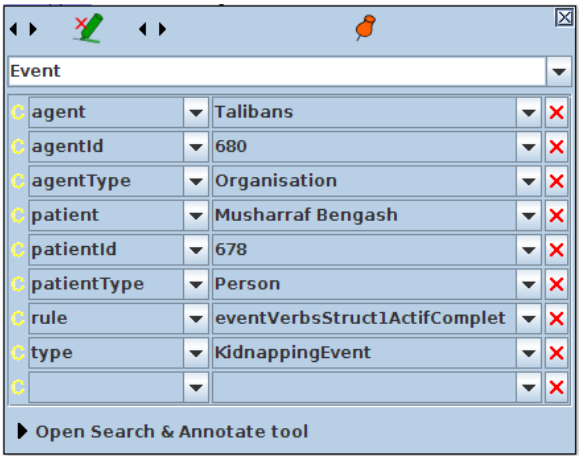


FIGURE 5.14 – Extraction des événements : exemple d'annotation GATE

5.4.2 Apprentissage de patrons linguistiques

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.

Présentation de l'approche

La découverte de motifs séquentiels a été introduite par [Agrawal et al., 1993] dans le domaine de la fouille de données et adaptée par [Béchet et al., 2012] à l'extraction d'information dans les textes.

Ceux-ci s'intéressent en particulier à la découverte de motifs séquentiels d'*itemsets*. Il s'agit de repérer, dans un ensemble de séquences de texte, 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. Un *itemset* est un ensemble d'*items* décrivant un élément de cette séquence. Un *item* correspond à une caractéristique particulière de cet élément. Un certain nombre de paramètres peuvent être adaptés selon l'application visée : la nature de la séquence et des *items*, le nombre d'*items*, le 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 :

```
< three Chileans arrested near Buenos Aires >
< NP arrested near Location >
< NP VB PRP Location >
```

où *NP* est un syntagme nominal, *VB* est un verbe, *PRP* est une préposition et *Location* est une entité nommée de type *Lieu*.

La phase d'apprentissage permet d'obtenir un ensemble de motifs séquentiels fréquents qui sont ensuite sélectionnés par un expert pour en retenir les plus pertinents pour la tâche d'extraction visée. Les motifs retenus sont alors appliqués sur un le nouveau corpus à analyser, préalablement annoté pour obtenir les différents types d'*items* considérés.

Contrairement à d'autres approches d'EI (présentées en section 2.2), la découverte de motifs séquentiels fréquents 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 aux performances inégales et peu disponible librement selon les langues. Toutefois, le point faible partagé par les méthodes d'apprentissage symbolique 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, 2007] pour ordonner et visualiser les motifs des plus généraux aux plus spécifiques puis filtrer les plus pertinents.

Application à l'extraction automatique des événements

Dans la lignée de ces travaux, nous avons utilisé un outil d'extraction de motifs séquentiels développé au GREYC [110] (selon la méthode de [Béchet et al., 2012]). Le système repris présente plusieurs points forts qui justifient ce choix : il permet d'extraire des motifs dits "fermés" (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 de fouille de données existants.

Notre contribution a donc été d'adapter la technique de fouille de motifs séquentiels à notre domaine d'application et au traitement de dépêches de presse dans le but de générer automatiquement des patrons linguistiques pour la détection d'événements. Nous avons tout d'abord défini un ensemble de paramètres de base pour l'apprentissage : nous choisissons comme séquence la phrase et comme unité de base le

110. https://sdmc.greyc.fr

token ainsi qu'un ensemble d'items de mot tels que sa forme fléchie, son lemme, sa catégorie grammaticale. Pour segmenter le corpus d'apprentissage et obtenir ces différentes informations, nous avons employé l'analyseur morpho-syntaxique TreeTagger [111] [Schmid, 1994]. A cela, nous proposons d'ajouter une reconnaissance des entités nommées (de type *Personne*, *Organisation*, *Lieu* et *Date*) ainsi qu'un repérage lexical des déclencheurs d'événements. Nous pouvons ainsi découvrir des motifs séquentiels fréquents impliquant un déclencheur d'événement et une ou plusieurs entités d'intérêt constituant les participants/circonstants de l'événement en question. Ces deux traitements sont réalisés grâce à la chaine de REN présentée en section 5.3 et aux *gazetteers* construits pour le premier système symbolique.

Enfin, pour réduire le nombre de motifs retournés et faciliter la sélection manuelle de l'expert, nous introduisons un ensemble de contraintes spécifiques à notre application. D'une part, des contraintes linguistiques d'appartenance permettant de filtrer les motifs selon les items qu'ils contiennent. Pour notre application, la contrainte d'appartenance utilisée est de ne retourner que les motifs contenant au minimum un déclencheur d'événement et une entité nommée d'un type d'intérêt (voir plus haut). D'autre part, nous avons employé une contrainte dite de *gap* [Dong and Pei, 2007], permettant l'extraction de motifs ne contenant pas nécessairement des *itemsets* consécutifs (contrairement aux n-grammes dont les éléments sont strictement contigus). Un *gap* d'une valeur maximale *n* signifie qu'au maximum *n itemsets* (mots) sont présents entre chaque *itemset* du motif retourné dans les séquences correspondantes. Par exemple, si la séquence suivante *(1)* est présente dans le corpus d'apprentissage et que le *gap* est fixé à 2, le système pourra retourner le motif *(2)* :

(1) [...] a suspected sectarian attack in Mehmoodabad area of Karachi [...]

(2) < DT EventTrigger PRP Location PRP Location >

où *DT* est un déterminant, *EventTrigger* un déclencheur d'événement, *PRP* une préposition et *Location* une entité nommée de type Lieu. La contrainte de *gap* permet, dans cet exemple, de retourner un motif plus générique en omettant la présence des mots "suspected" et "sectarian" si ceux-ci ne sont pas fréquents. Les contraintes de *gap* et de support (fréquence minimale d'un motif) sont des paramètres à ajuster lors de la phase d'apprentissage. Ce paramétrage a été réalisé pour nos expérimentations et est présenté en partie 7.2.1.

Une fois le nombre de motifs extraits diminué grâce à des contraintes, ceux-ci ont été manuellement sélectionnés en fonction de leur pertinence pour la tâche d'extraction des événements. Pour cela, nous utilisons l'outil Camelis [Ferré, 2007] permettant d'ordonner et visualiser les motifs des plus généraux aux plus spécifiques puis de filtrer les plus pertinents. La figure 5.15 présente un aperçu de cet outil.

Les motifs ainsi sélectionnés sont ensuite appliqués sur un nouveau corpus afin d'en extraire les événements-cibles. Dans un souci de réutilisation des systèmes déjà développés et pour faciliter cette dernière étape, nous choisissons d'exprimer les motifs obtenus grâce au formalisme JAPE et obtenons ainsi un nouveau module intégrable dans une chaine de traitement GATE.

5.5 Conclusions

Notre seconde contribution détaillée dans ce chapitre est une approche mixte pour l'extraction automatique des événements fondée sur une méthode symbolique à base de grammaires contextuelles et sur

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

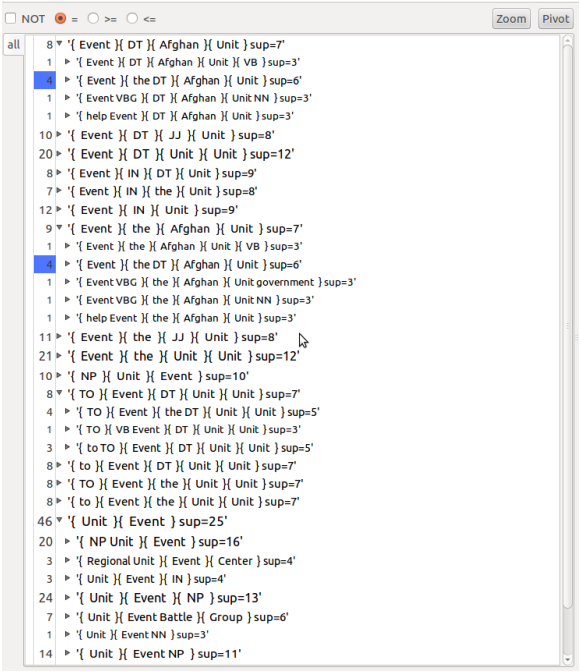


FIGURE 5.15 – Visualisation et sélection des motifs avec l'outil Camelis

une seconde technique de fouille de motifs séquentiels fréquents. Dans un premier temps, nous avons implémenté un extracteur d'entités nommées sur le modèle des approches symboliques classiques. Celui-ci permet d'annoter au préalable les différentes entités dites simples nécessaires à la reconnaissance des événements. Les résultats de ce premier extracteur ont montré des performances comparables à l'état de l'art bien qu'il pourrait être amélioré en réalisant notamment une extraction des dates dites relatives ou encore des entités d'intérêt autres que les entités nommées (comme l'extraction des équipements par exemple). Les deux méthodes pour l'extraction des événements ont montré leur efficacité lors de l'état de l'art présenté à la section 2.2.3 et nous renvoyons à la section 7.2 pour une évaluation de leurs performances sur un corpus de test de notre domaine d'application. La méthode à base de règles GATE pourra être améliorée en tenant compte d'autres informations fournies par l'analyse syntaxique telles que la voix (passive ou active) du déclencheur, la polarité de la phrase (négative ou positive), la modalité mais aussi les phénomènes de valence multiple. L'approche à base de motifs séquentiels fréquents pourrait également tirer profit de cette analyse syntaxique en intégrant les relations de dépendance produites en tant que nouveaux items ou sous forme de contraintes. Enfin, concernant les deux approches, leur limite principale (qui est aussi celle de beaucoup des approches de la littérature) est qu'ils réalisent l'extraction au niveau phrastique. Une granularité plus large tel que le paragraphe ou le discours pourrait permettre d'améliorer le rappel de ces approches.

Chapitre 6

Agrégation sémantique des événements

Sommaire

6.1 Introduction

L'état de l'art réalisé ainsi que le développement de notre propre système d'extraction d'information a montré la difficulté de cette tâche mais également et surtout des pistes d'amélioration possibles. Un premier constat est que les outils d'EI existants fournissent des résultats pouvant être incomplets, redondants, flous, conflictuels, imprécis et parfois totalement erronés. Par ailleurs, comme confirmé par nos premières expérimentations (voir la section 7.2), plusieurs approches actuelles s'avèrent complémentaires et une combinaison adaptée de leurs résultats pourrait aboutir à une découverte de l'information plus efficace. Dans le cadre de cette thèse, l'objectif visé est d'améliorer la qualité des fiches de connaissances générées et présentées à l'utilisateur afin de faciliter son processus de découverte des informations d'intérêt. Pour ce faire, nous proposons un processus d'agrégation sémantique des résultats issus de la phase d'extraction des événements. Cette troisième contribution vise à produire un ensemble de connaissances cohérent [112] en regroupant entre elles les mentions d'événements référant à un seul et même événement du monde réel. Ce processus est fondé sur des calculs de similarité sémantique et permet d'agréger à la fois des événements provenant des deux extracteurs développés et/ou de différentes sources d'information (agrégation intra- et inter-documents).

Dans ce chapitre, nous présentons tout d'abord les mécanismes mis en place pour la normalisation des entités extraites, étape préalable nécessaire au processus d'agrégation. Dans un second temps, sont exposées les différentes méthodes définies pour estimer la similarité des événements au niveau de chacune de leurs dimensions (conceptuelle, temporelle, spatiale et agentive). Enfin, nous détaillons le processus d'agrégation proposé, fondé sur ces mesures de similarité locales.

6.2 Normalisation des entités

Notre processus d'agrégation sémantique des événements nécessite de manipuler non plus de simples portions de texte extraites mais des objets sémantiques, des éléments de connaissance à proprement parler. Pour cela, nous proposons une première phase de normalisation visant à désambiguïser les différentes entités impliquées dans les événements dans le but de les rattacher à un objet du monde réel (c'est-à-dire déterminer leur référence). Nous présentons, dans les sections suivantes, les différentes méthodes mises en place pour la normalisation de ces entités.

Normalisation des dates

La normalisation des entités temporelles est nécessaire en raison des multiples façons possibles d'exprimer le temps en langage naturel ("04/09/2012", "Mon, 09/April/12", "two days ago", etc.). Celle-ci consiste à les convertir en un format unique facilitant leur comparaison et le calcul de leur similarité (voir la section 6.3.2). Conformément à notre modélisation temporelle de l'événement (voir la section 4.2.2), nous avons choisi pour cela le format TUS, définissant une représentation numérique unifiée des dates. Nous avons effectué cette normalisation au sein-même de notre module d'extraction d'information : les règles JAPE en question récupèrent séparément les différents éléments de la date (année, mois et jour,

112. Nous entendons ici par "cohérence" une absence de contradictions et de redondances.

les granules dans le modèle TUS) afin de reconstituer sa forme normalisée et d'ajouter cette information sous forme d'attribut aux annotations de type *Date* produites.

Nous avons fait le choix de nous intéresser exclusivement aux dates absolues car la résolution des dates relatives est un sujet de recherche à part entière [Llorens Martínez, 2011] que nous ne pouvons approfondir dans le cadre de cette thèse. En effet, l'extraction de ces dernières nécessite des traitements plus complexes en vue de leur normalisation tels que la résolution d'anaphores temporelles, l'analyse automatique du contexte linguistique d'occurrence, etc.

Le tableau 6.1 ci-dessous présente quelques exemples de dates extraites dans des textes en anglais ainsi que leurs formes normalisées par notre outil.

Date extraite	Date normalisée
1999-03-12	1999-03-12
1948	1948-01-01/12-31
April 4, 1949	1949-04-04
July 1997	1997-07-01/31
03-12-99	1999-12-03

TABLE 6.1 – Normalisation des dates

Normalisation des lieux

[Garbin and Mani, 2005] ont montré que 40% des toponymes présents dans un corpus de dépêches de l'AFP sont ambigus. Il existe différents types d'ambiguïté : un nom propre peut référer à des entités de types variés (*Paris* peut faire référence à une ville ou à une personne), deux lieux géographiques peuvent porter le même nom (*London* est une ville en Angleterre mais aussi au Canada), certaines entités géographiques peuvent également porter plusieurs noms ou en changer au cours du temps, etc.

Notre objectif ici est d'associer à chaque entité géographique extraite un référent unique correspondant à une entité du monde réel bien définie. Nous avons pour cela utilisé un service de désambiguïsation des entités géographiques développé dans notre département. Celui-ci opère en attribuant à chaque entité géographique extraite un identifiant de la base sémantique GeoNames. Le projet GeoNames a notamment pour avantages d'être *open-source*, de définir de nombreux toponymes dans le monde (l'information n'est pas limitée à certaines régions géographiques) et d'intégrer des relations avec d'autres sources de données (permettant des traitements plus avancés si nécessaire). Par ailleurs, ce composant de normalisation des lieux s'inspire des travaux de [Buscaldi, 2010]. L'approche est fondée sur des calculs de cohérence sur l'ensemble des toponymes possibles pour chaque entité géographique extraite. Cette cohérence est définie à partir de trois types de critères :

– les distances géographiques entre toponymes ;
– les types géographiques des toponymes (continent, ville, rivière ...) ;
– les distances entre les entités géographiques au sein du texte considéré.

L'intérêt de cette méthode est de pouvoir dégager des groupes de toponymes cohérents entre eux. De plus, elle permet de donner plus d'importance à la notion de qualité administrative par rapport à celle de proximité géographique lorsqu'une dépêche parle, par exemple, de *New York*, *Paris* et *Londres*.

Le code ci-dessous est un extrait des triplets RDF/XML produits par ce service. Les quatre premières lignes lient l'entité extraite identifiée par l'URI *http ://weblab.ow2.org/wookie/instances/Place#india* et son entité de référence dans la base GeoNames (ayant pour identifiant *http ://sws.geonames.org/1269750/*). Le reste de l'exemple donne quelques-unes des propriétés de l'entité GeoNames *Republic of India* : sa classe dans l'ontologie Geonames (ligne 15), ses coordonnées géographiques selon le standard WGS84 (lignes 9 et 11), son nom dans GeoNames (ligne 14), etc.

```
1  <rdf:Description
2   rdf:about="http://weblab.ow2.org/wookie/instances/Place#india">
3          <geo:hasSpatialThing rdf:resource="http://sws.geonames.org/1269750/"/>
4  </rdf:Description>
5
6  <rdf:Description rdf:about="http://sws.geonames.org/1269750/">
7          <geo:parentFeature rdf:resource="http://sws.geonames.org/6295630/"/>
8          <geo:parentFeature rdf:resource="http://sws.geonames.org/6255147/"/>
9          <wgs84:long
10         rdf:datatype="http://www.w3.org/2001/XMLSchema#double">77.0</wgs84:long>
11         <wgs84:lat
12         rdf:datatype="http://www.w3.org/2001/XMLSchema#double">20.0</wgs84:lat>
13         <rdfs:label>Republic of India</rdfs:label>
14         <geo:name>Republic of India</geo:name>
15         <rdf:type rdf:resource="http://www.geonames.org/ontology#Feature"/>
16 </rdf:Description>
```

FIGURE 6.1 – Désambiguïsation des entités spatiales : exemple de triplets RDF/XML produits

Normalisation des participants

La normalisation des participants est également une problématique complexe de par la diversité des noms de personnes et d'organisations rencontrés dans les dépêches de presse. Une même entité est souvent mentionnée sous de multiples formes telles que des abréviations (Boston Symphony Orchestra vs. BSO), des formes raccourcies (Osama Bin Laden vs. Bin Laden), des orthographes alternatives (Osama vs. Ussamah vs. Oussama), des alias ou pseudonymes (Osama Bin Laden vs. Sheikh Al-Mujahid). De nombreux travaux se sont intéressés à la désambiguïsation de telles entités et particulièrement depuis l'essor du mouvement *Linked Data* et la mise à disposition de nombreuses bases de connaissances sémantiques sur le Web (voir le chapitre 3). Bien que cette thèse ne nous ait pas permis d'approfondir cette problématique, nous avons retenu, à titre de perspectives, certaines approches théoriques et outils qui pourront être intégrés au sein de notre processus d'agrégation.

Nous pouvons citer, parmi les outils disponibles sur le Web, DBPedia Spotlight [113] et Zemanta [114], permettant l'attribution d'un référent unique (provenant d'une base de connaissances externe) à chaque entité d'intérêt repérée dans un texte. Le premier outil est issu d'un projet *open source* et se présente sous la forme d'un service Web utilisant DBPedia comme base sémantique de référence. Le système Zemanta est lui propriétaire et initialement conçu pour aider les créateurs de contenu Web (éditeurs, blogueurs, etc.) grâce à des suggestions automatiques diverses (images, liens, mots-clés, etc.). Cet outil suggère notamment des liens sémantiques vers des concepts issus de nombreuses bases de connaissances telles que Wikipédia, IMDB, MusicBrainz, etc. Du côté des travaux théoriques, nous avons retenu deux

113. https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki
114. http://developer.zemanta.com/

approches qui nous paraissent adaptées pour réaliser cette étape de normalisation des participants d'un événement. D'une part, [Cucerzan, 2007] propose de réaliser de façon conjointe la reconnaissance des entités nommées et leur désambiguïsation. Pour cette seconde tâche, cette approche emploie une grande quantité d'informations contextuelles et catégorielles automatiquement extraites de la base Wikipédia. Il est notamment fait usage des nombreux liens existants entre les entités de cette base, des pages de désambiguïsation sémantique (redirections) ainsi que d'un modèle vectoriel créé à partir du contexte textuel des entités nommées repérées. Par ailleurs, [Mann and Yarowsky, 2003] présente un ensemble d'algorithmes visant à déterminer le bon référent pour des noms de personne en utilisant des techniques de regroupement non-supervisé et d'extraction automatique de caractéristiques. Les auteurs montrent notamment comment apprendre et utiliser automatiquement l'information biographique contenue dans les textes (date de naissance, profession, affiliation, etc.) afin d'améliorer les résultats du regroupement. Cette technique a montré de bons résultats sur une tâche de désambiguïsation de pseudonymes.

6.3 Similarité sémantique entre événements

Nous proposons d'évaluer la similarité entre événements, c'est-à-dire d'estimer à quel degré deux mentions d'événement peuvent référer à un seul et même événement de la réalité. Cette estimation sera utilisée dans notre application pour aider l'utilisateur final à compléter ses connaissances et à prendre des décisions de fusion d'information. Celui-ci pourra, le cas échéant, décider de fusionner deux fiches de connaissances (une interface d'aide à la fusion de fiches a été développée au sein de la plateforme WebLab) qu'il considère référer au même événement réel. Ci-après, nous détaillons les mesures de similarité mises en œuvre pour chacune des dimensions de l'événement et les exprimons selon une échelle qualitative. En effet, une échelle qualitative s'avère plus pertinente pour élaborer un processus de capitalisation orienté utilisateur : étant donné que ces similarités seront présentées à l'analyste, une estimation symbolique sera mieux comprise qu'une similarité numérique. Nous définissons cette échelle comme composée de quatre niveaux :

1. identité (ID) : les deux dimensions réfèrent à la même entité réelle,
2. proximité (PROX) : les deux dimensions ont des caractéristiques communes et pourraient référer à la même entité du monde réel,
3. différence (DIFF) : les deux dimensions ne réfèrent pas à la même entité,
4. manque d'information (MI) : il y a un manque d'information dans l'une ou les deux dimensions (résultant du document d'origine ou du système d'extraction) qui empêche de savoir si elles réfèrent ou non à la même entité réelle.

Définition 4. *Une fonction de similarité sémantique est une fonction :*

$$R : \mathcal{E} \times \mathcal{E} \rightarrow \{ID, PROX, DIFF, MI\}$$

où $\mathcal{E}$ est un ensemble d'événements et ID, PROX, DIFF et MI représentent respectivement l'identité, la proximité, la différence et le manque d'information.

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

6.3.1 Similarité conceptuelle

Soient $e = (S, T, SP, A)$ et $e' = (S', T', SP', A')$ deux événements et S et S' leurs dimensions conceptuelles respectives.

Définition 5. *R_s est une fonction de similarité conceptuelle :*

$$R_s : \mathcal{E} \times \mathcal{E} \rightarrow \{ID, PROX, DIFF\}$$

tel que

1. $R_s(e, e') = DIFF$ *ssi* S *et* S' *sont deux classes différentes de l'ontologie et ne sont pas en relation de subsomption l'une avec l'autre*
2. $R_s(e, e') = PROX$ *ssi* S *est une sous-classe (à tout niveau) de* S' *dans l'ontologie (et inversement)*
3. $R_s(e, e') = ID$ *ssi* S *et* S' *correspondent à la même classe de l'ontologie de domaine*

Notons ici que le niveau *Manque d'Information (MI)* n'est pas défini car la dimension conceptuelle d'un événement est nécessairement fournie par notre système d'extraction. En effet, celui-ci est conçu de façon telle qu'il n'y a pas d'extraction d'événement sans déclencheur d'événement et association à une classe de l'ontologie.

Prenons quelques exemples pour illustrer le calcul de cette similarité conceptuelle.

Exemple 2. *Soient e_1, e_2, e_3 et e_4 des événements et S_1, S_2, S_3 et S_4 leurs dimensions conceptuelles respectives tel que $S_1 = AttackEvent$, $S_2 = BombingEvent$, $S_3 = SearchOperation$ et $S_4 = AttackEvent$.*

1. $R_s(e_1, e_3) = DIFF$ *car les classes S_1 et S_3 ne sont pas en relation de subsomption dans notre ontologie de domaine (voir l'annexe B).*
2. $R_s(e_1, e_2) = PROX$ *car S_2 est une sous-classe de S_1 dans l'ontologie (voir l'annexe B).*
3. $R_s(e_1, e_4) = ID$ *car e_1 et e_4 sont des instances de la même classe d'événement dans WOOKIE ($S_1 = S_4 = AttackEvent$).*

6.3.2 Similarité temporelle

Soient $T = [g_1, g_2, g_3]$ et $T' = [g'_1, g'_2, g'_3]$ deux entités temporelles exprimées au formalisme TUS.

Définition 6. *Nous définissons $\oplus$, un opérateur de complétion temporelle, prenant en paramètres T et T' où T et/ou T' est incomplet (voir la section 4.2.2) et retourne une expression temporelle $T'' = [g''_1, g''_2, g''_3]$ où :*

1. *si $g_i = g'_i$ alors $g''_i = g_i = g'_i$*
2. *si $g_i = \emptyset$ alors $g''_i = g'_i$ (et inversement)*

Exemple 3. *Si $T = [\emptyset, 3, 15]$ et $T' = [2012, \emptyset, \emptyset]$ alors $T'' = T \oplus T' = [2012, 3, 15]$*

Exemple 4. *Si $T = [\emptyset, 2, \emptyset]$ et $T' = [2013, \emptyset, 29]$ alors $T'' = T \oplus T' = [2013, 2, 29]$*

Dans certains cas, deux événements ayant deux dates d'occurrence strictement différentes selon le format TUS (niveau *DIFF* défini plus bas) peuvent tout de même référer au même événement dans le monde réel. Par exemple, une attaque peut durer plusieurs heures sur deux jours et ce même événement pourra être reporté avec deux dates différentes. Afin de traiter ces situations particulières, nous définissons $D_t(T, T')$, une fonction générique retournant une distance temporelle entre deux entités et un seuil d_t délimitant la différence (*DIFF*) et la proximité (*PROX*) temporelle. Cette distance peut être obtenue par des librairies dédiées à la manipulation d'entités temporelles et le seuil d_t défini en fonction de l'application.

Soient $e = (S, T, SP, A)$ et $e' = (S', T', SP', A')$ deux événements, nous définissons R_t au moyen de ces différents opérateurs et seuils.

Définition 7. *R_t est une fonction de similarité temporelle :*

$$R_t : \mathcal{E} \times \mathcal{E} \rightarrow \{ID, PROX, DIFF, MI\}$$

telle que

1. $R_t(e, e') = DIFF$ *ssi l'une des conditions suivantes est vérifiée :*

$$\begin{cases} \exists_{i \in \{1,2,3\}} \text{ tel que } g_i \neq \emptyset \text{ et } g'_i \neq \emptyset \text{ et } g_i \neq g'_i \text{ et } D_t(T, T') > d_t \\ \text{si } T \text{ ou } T' \text{ est incomplet et } T \oplus T' \text{ retourne une date illégale (voir la section 4.2.2)} \end{cases}$$

2. $R_t(e, e') = PROX$ *ssi l'une des conditions suivantes est vérifiée :*

$$\begin{cases} R_t(e, e') = DIFF \text{ et } D_t(T, T') \leq d_t \\ T \text{ ou } T' \text{ est incomplet et } T \oplus T' \text{ retourne une date légale (voir la section 4.2.2)} \end{cases}$$

3. $R_t(e, e') = ID$ *ssi* T *et* T' *sont complets (voir la section 4.2.2) et* $\forall_{i \in \{1,2,3\}}, g_i = g'_i$
4. $R_t(e, e') = MI$ *ssi* T *et/ou* T' *est inconnu (c'est-à-dire que soit l'information n'était pas présente dans la source, soit elle n'a pas été extraite par le système d'extraction)*

Prenons quelques exemples pour illustrer le calcul de cette similarité temporelle.

Exemple 5. *Soient e_1, e_2, e_3, e_4 et e_5 des événements et T_1, T_2, T_3, T_4 et T_5 leurs dimensions temporelles respectives tel que $T_1 = [2012, 11, 05]$, $T_2 = [2013, 11, 05]$, $T_3 = \emptyset$, $T_4 = [2012, 11, \emptyset]$ et $T_5 = [2013, 11, 05]$.*

1. $R_t(e_1, e_2) = DIFF$
2. $R_t(e_1, e_4) = PROX$
3. $R_t(e_2, e_5) = ID$
4. $R_t(e_2, e_3) = MI$

6.3.3 Similarité spatiale

Pour estimer la similarité entre entités géographiques, nous utilisons les relations topologiques du modèle RCC-8 (présenté en section 4.2.3) ainsi que les différentes relations spatiales existantes dans la base GeoNames. Comme mentionné précédemment, la dimension spatiale peut varier avec le point de

vue duquel un événement est rapporté : par exemple, un événement survenu à Cestas (un petit village à côté de Bordeaux) pourra être précisément situé par une agence de presse française mais pourrait être plus généralement localisé à Bordeaux (la grande ville la plus proche) par une agence étrangère. La similarité peut également être influencée par la nature des deux entités spatiales comparées (ville, pays, quartier, continent, etc.) et par la taille de la région administrative les englobant (par exemple, la distance entre deux villes du Lichtenstein ne sera pas considérée de la même façon que celle entre deux villes de Russie). Pour résumer, l'absence d'identité topologique entre deux lieux peut être due à une différence dans le niveau d'abstraction mais ne signifie pas nécessairement une différence spatiale (niveau *DIFF*). Pour mieux traiter ces cas, nos introduisons $D_{sp}(SP, SP')$, une fonction générique retournant une distance spatiale entre deux entités géographiques et un seuil d_{sp} délimitant la différence (*DIFF*) et la proximité (*PROX*) spatiale. Cette distance peut être obtenue par des librairies dédiées à la manipulation d'entités spatiales et le seuil d_{sp} défini en fonction de l'application.

Soient $e = (S, T, SP, A)$ et $e' = (S', T', SP', A')$ deux événements et r une relation RCC-8 entre deux entités spatiales SP et SP' (voir la section 1.3.2 et la figure 1.11).

Définition 8. *R_{sp} est une fonction de similarité spatiale :*

$$R_{sp} : \mathcal{E} \times \mathcal{E} \rightarrow \{ID, PROX, DIFF, MI\}$$

telle que

1. $R_{sp}(e, e') = DIFF$ *ssi* $r(SP, SP') = DC$ *et* $D_{sp}(SP, SP') > d_{sp}$
2. $R_{sp}(e, e') = PROX$ *ssi l'une des conditions suivantes est vérifiée :*

$$\begin{cases} r(SP, SP') \in \{EC, PO, TPP, NTPP, TPPi, NTPPi\} \\ r(SP, SP') = DC \text{ et } D_{sp}(SP, SP') < d_{sp} \end{cases}$$

3. $R_{sp}(e, e') = ID$ *ssi* $r(SP, SP') = EQ$
4. $R_{sp}(e, e') = MI$ *ssi* SP *et/ou* SP' *est inconnu (c'est-à-dire que soit l'information n'était pas présente dans la source, soit elle n'a pas été extraite par le système d'extraction)*

Ces définitions théoriques sont implémentées en utilisant les relations GeoNames comme suit :

Définition 9.
1. $R_{sp}(e, e') = DIFF$ *ssi* SP *et* SP' *ne sont liés par aucune relation topologique dans GeoNames et* $D_{sp}(SP, SP') > d_{sp}$
2. $R_{sp}(e, e') = PROX$ *ssi*
   (a) SP *et* SP' *sont liés par une relation "nearby", "neighbour" ou "children" dans la base GeoNames*
   (b) SP *et* SP' *ne sont liés par aucune relation topologique dans GeoNames et* $D_{sp}(SP, SP') < d_{sp}$
3. $R_{sp}(e, e') = ID$ *ssi* SP *et* SP' *ont le même identifiant GeoNames (c'est à dire la même URI)*
4. $R_{sp}(e, e') = MI$ *ssi* SP *et/ou* SP' *est inconnu (c'est-à-dire que soit l'information n'était pas présente dans la source, soit elle n'a pas été extraite par le système d'extraction)*

Prenons quelques exemples pour illustrer le calcul de cette similarité spatiale.

Exemple 6. *Soient e_1, e_2, e_3, e_4 et e_5 des événements et SP_1, SP_2, SP_3, SP_4 et SP_5 leurs dimensions spatiales respectives tel que $SP_1 = Paris$, $SP_2 = Singapour$, $SP_3 = \emptyset$, $SP_4 = France$ et $SP_5 = Singapour$.*

1. $R_sp(e_1, e_2) = DIFF$
2. $R_sp(e_1, e_4) = PROX$
3. $R_sp(e_2, e_5) = ID$
4. $R_sp(e_2, e_3) = MI$

6.3.4 Similarité agentive

Comme mentionné précédemment, l'ensemble des participants d'un événement est composé d'entités nommées de type *Personne* ou *Organisation* et ces participants sont concrètement stockés dans une base de connaissances sous la forme de chaines de caractères. Par conséquent, l'agrégation au niveau de cette dimension implique l'utilisation de mesures de similarité dédiées aux chaines de caractères mais aussi adaptées à la comparaison des entités nommées. La distance de Jaro-Winkler [Winkler et al., 2006] convient bien à notre cadre d'application en raison de son temps d'exécution modéré et sa bonne performance dans le traitement des similarités entre noms de personne notamment. Notre approche visant à rester indépendante des mesures de similarité choisies, il conviendra d'évaluer plusieurs métriques au sein de notre système de capitalisation des connaissances pour guider notre choix final et définir le seuil de distance d_{str} utilisé ci-dessous.

Nous définissons $D_{str}(P, P')$, une fonction générique retournant une distance entre deux chaines de caractères tel que $D_{str}(P, P') = 0$ signifie que les chaines de caractères représentant les participants P et P' sont égales.

De plus, comme nous manipulons des entités nommées et non de simples chaines de caractères, nous avons exploré d'autres techniques dédiées à la résolution de coréférence entre entités nommées telles que [Finin et al., 2009]. Dans un premier temps, nous proposons d'utiliser la base sémantique DBpedia [115] pour agréger tous les noms alternatifs référant à la même entité réelle. Nous définissons $alt(P)$ comme une fonction retournant tous les noms alternatifs donnés dans DBpedia pour une participant P.

Définition 10. *P et P' coréfèrent, noté $P \simeq P'$, ssi $P \in alt(P')$ (et inversement) ou $D_{str}(P, P') = 0$*

Par ailleurs, comme mentionné en section 4.2.4, la dimension A est définie comme un ensemble de participants, par conséquent, l'agrégation des participants d'un événement doit permettre de traiter une similarité entre ensembles.

Définition 11. *A et A' coréfèrent, noté $A \cong A'$, ssi $|A| = |A'|$ et $\forall P \in A$, $\exists P' \in A'$ tel que $P \simeq P'$ (et inversement)*

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 $A' = \{P'_1, P'_2, \ldots, P'_n\}$.

115. http://dbpedia.org/

Définition 12. *R_a est une fonction de similarité agentive :*

$$R_a : \mathcal{E} \times \mathcal{E} \rightarrow \{ID, PROX, DIFF, MI\}$$

tel que

1. $R_a(e, e') = DIFF$ *ssi* $\forall_{P \in A}$ *et* $\forall_{P' \in A'}$, $P \notin alt(P')$ *(et inversement) et* $D_{str}(P, P') > d_{str}$
2. $R_a(e, e') = PROX$ *ssi l'une des conditions suivantes est vérifiée :*

$$\begin{cases} \forall_{P \in A}, \exists_{P' \in A'} \text{ tel que } P \simeq P' \\ \forall_{P \in A}, \exists_{P' \in A'} \text{ tel que } D_{str}(P, P') \leq d_{str} \end{cases}$$

3. $R_a(e, e') = ID$ *ssi* $A \cong A'$
4. $R_a(e, e') = MI$ *ssi* $|A| = 0$ *et/ou* $|A'| = 0$ *(c'est-à-dire que soit l'information n'était pas présente dans la source, soit elle n'a pas été extraite par le système d'extraction)*

Prenons quelques exemples pour illustrer le calcul de cette similarité agentive.

Exemple 7. *Soient* e_1, e_2, e_3, e_4 *et* e_5 *des événements et* A_1, A_2, A_3, A_4 *et* A_5 *leurs dimensions agentives respectives tel que* $A_1 = \{$*Elmer Eusebio*$\}$, $A_2 = \{$*Police*$\}$, $A_3 = \{\emptyset\}$, $A_4 = \{$*E. Eusebio*$\}$ *et* $A_5 = \{$*Police*$\}$.

1. $R_a(e_1, e_2) = DIFF$
2. $R_a(e_1, e_4) = PROX$
3. $R_a(e_2, e_5) = ID$
4. $R_a(e_2, e_3) = MI$

6.4 Processus d'agrégation

Nous proposons un processus d'agrégation fondé sur une représentation en graphe : l'ensemble des similarités calculées est organisé en un graphe non-orienté $G = (E, edges)$, où E est l'ensemble des sommets et chaque sommet $e \in E$ est un événement extrait, et où $edges$ est l'ensemble des arêtes et chaque arête $edge$ est un lien entre deux sommets e et $e\prime$ défini comme une fonction de similarité multivaluée.

Définition 13. *La fonction* $edge$ *a pour domaine de définition :*

$$edge : \mathbf{E} \times \mathbf{E} \rightarrow \mathbf{R}$$

où $\mathbf{R} = R_s \times R_t \times R_{sp} \times R_a$ *correspondant aux fonctions de similarité définies précédemment.*

Définition 14. *La fonction* $edge$ *est définie telle que :*

$$edge(e, e\prime) = \langle R_s(e, e\prime), R_t(e, e\prime), R_{sp}(e, e\prime), R_a(e, e\prime) \rangle$$

Cette représentation est bien adaptée à notre cadre d'application (la plateforme WebLab et les technologies du Web sémantique au sens large), car la connaissance provenant des différents services de traitement de l'information est stockée dans des bases de connaissances sémantiques fondées sur les graphes.

Le graphe G est construit selon le principe suivant : chaque événement extrait (avec l'ensemble de ses dimensions) est comparé deux à deux à chacun des autres événements pour déduire un degré de similarité au niveau de chaque dimension (R_s, R_t, R_{sp} et R_a) (selon les principes théoriques exposés dans les sections précédentes). Nous créons ensuite dans G l'ensemble des sommets E correspondant aux événements extraits ainsi qu'une arête multivaluée de similarité $edge$ entre chaque paire d'événements.

Les similarités obtenues pour chacune des dimensions (conceptuelle, temporelle, spatiale et agentive) constituent un ensemble d'indicateurs que nous souhaitons exploiter afin de déterminer si des événements co-réfèrent. 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 (sa véracité, sa fiabilité, etc.), 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 toute perte d'information pouvant survenir avec une fusion entièrement automatisée. L'objectif de nos travaux est en effet, non pas de concevoir un système de fusion de fiches à proprement parler, mais plutôt d'aider l'analyste dans sa prise de décision grâce à un processus d'agrégation sémantique des événements.

Pour ce faire, nous proposons d'appliquer un regroupement automatique (*clustering*) sur le graphe de similarités obtenu afin de guider l'analyste au sein de cet ensemble de connaissances. Différentes combinaisons de similarités sont possibles pour réaliser ce regroupement en fonction des besoins de l'utilisateur. Nous proposons, dans un premier temps, de hiérarchiser l'ensemble de ces configurations selon leur degré de similarité global (en commençant par celle qui lie les événements les plus similaires) de la manière suivante :

Configuration $(C1)$ $\{isConceptuallyID, isTemporallyID, isSpatiallyID, isAgentivelyID\}$
Configuration $(C2a)$ $\{isConceptuallyID, isTemporallyID, isSpatiallyID, isAgentivelyPROX\}$
Configuration $(C2b)$ $\{isConceptuallyID, isTemporallyID, isSpatiallyPROX, isAgentivelyID\}$
Configuration $(C2c)$ $\{isConceptuallyID, isTemporallyPROX, isSpatiallyID, isAgentivelyID\}$
Configuration $(C2d)$ $\{isConceptuallyPROX, isTemporallyID, isSpatiallyID, isAgentivelyID\}$
Configuration $(C3a)$ $\{isConceptuallyID, isTemporallyID, isSpatiallyPROX, isAgentivelyPROX\}$
Configuration (Cn) etc.

Intuitivement, la configuration $C1$ parait la plus à même de regrouper entre eux les événements qui co-réfèrent. Toutefois, l'incomplétude et l'hétérogénéité des données extraites est telle que cette première configuration est peu fréquemment observée dans une application réelle. De sorte que, si l'analyste souhaite retrouver dans sa base de connaissances des événements similaires, il sera plus pertinent d'explorer d'autres configurations d'agrégation. Pour cela, nous proposons de réaliser un regroupement ascendant et hiérarchique du graphe de similarités fondé sur les différentes configurations de similarités possibles.

Le processus de regroupement se déroule ainsi :

1. Nous appliquons une première passe de regroupement selon la première configuration de notre hiérarchie $C1$ afin d'obtenir un premier jeu de n agrégats d'événements Ω.
   Ainsi, chaque sommet $e \in \Omega_i$ est lié à tous les autres sommets de Ω_i par une arête $edge_{C1}$ satisfaisant la configuration $C1$.
   Par ailleurs, chaque agrégat Ω_i ainsi formé possède la caractéristique suivante : l'ensemble des arêtes $edgeCn$ (présentes avant le premier regroupement) reliant un sommet $e \in \Omega_i$ à un autre sommet $e\prime \in \Omega_j$ sera de même configuration de similarités.
2. Nous pouvons donc fusionner cet ensemble d'arêtes en une nouvelle arête $edgeCx$ où Cx correspond à cette configuration commune et relie l'agrégat Ω_i au sommet $e \in \Omega_j$.
3. Une fois ce nouvel ensemble d'arêtes créé, nous proposons d'effectuer de nouveau un regroupement selon une nouvelle configuration Cx (la suivante dans la hiérarchie) et ainsi de suite.

Ce processus de regroupements successifs n'a pas vocation à être réalisé jusqu'à épuisement des configurations possibles. En effet, afin de proposer un environnement d'aide à la décision flexible et adaptable, nous envisageons de laisser l'analyste libre du nombre de regroupements qu'il souhaite effectuer ainsi que de définir ses propres configurations de regroupement. Cela sera réalisé de la façon suivante : les différents agrégats d'événements constitués après chaque phase de regroupement seront rendus accessibles à l'utilisateur au travers de diverses interfaces de la plateforme WebLab (recherche, carte géographique, bandeau temporel et autres vues). Celui-ci pourra observer les différents agrégats proposés et décider en fonction de cela du prochain regroupement à effectuer. Par exemple, l'utilisateur pourra demander au système de lui renvoyer tous les événements jugés similaires seulement sur le plan temporel (niveau *PROX*) et ensuite examiner les fiches d'événements correspondantes pour en déduire de nouvelles connaissances. Ce processus d'agrégation itératif et interactif sera illustré dans la section 7.3 de nos expérimentations.

6.5 Conclusions

Ce chapitre nous a permis de présenter un processus d'agrégation sémantique des événements fondé sur une échelle de similarité qualitative et sur un ensemble de mesures spécifiques à chaque type de dimension. Nous avons tout d'abord proposé des mécanismes de normalisation des entités, adaptés à leurs natures, afin d'harmoniser formellement les différentes informations extraites. Concernant ce premier aspect, nous envisageons des améliorations futures telles que la désambiguïsation des dates relatives, par exemple, ou encore l'intégration au sein de notre système d'un outil de désambiguïsation des participants (tels que ceux mentionnés en section 6.2). Nous avons ensuite proposé un une échelle de similarité qualitative orientée utilisateur et un ensemble de calculs de similarité intégrant à la fois un modèle théorique adapté à chaque dimension et une implémentation technique employant les technologies du Web sémantique (ontologies et bases de connaissances externes). Nous souhaitons poursuivre ce travail en élargissant le panel de similarités employées comme en intégrant des mesures de proximité ontologique plus sophistiquées ainsi des outils de distance temporelle et spatiale (fondés sur les coordonnées géographiques par exemple) et pour la similarité agentive, une distance dédiée aux ensembles telle que SoftJaccard [Largeron et al., 2009]. L'ensemble des similarités calculées pourraient également provenir d'une fonction d'équivalence apprise automatiquement à partir de données annotées manuellement. Enfin, un processus d'agrégation fondé sur les graphes a été proposé afin de regrouper les mentions d'événements similaires et de permettre à l'analyste de découvrir de nouvelles connaissances. Ce type d'agrégation possède l'avantage principal d'être intrinsèquement adapté aux traitement des bases de connaissances et

ainsi aisément généralisable à d'autres silos du Web de données. Cette agrégation pourrait également être réalisée par calcul d'une similarité globale qui combinerait les différentes similarités locales. Le point délicat pour cette méthode sera alors d'estimer le poids de chaque dimension dans cette combinaison. Enfin, la possibilité donnée à l'utilisateur de fusionner des fiches grâce aux suggestions d'agrégats soulèvera d'autres problématiques à explorer telles que la mise à jour et le maintien de cohérence de la base de connaissance en fonction des actions de l'analyste.

Chapitre 7

Expérimentations et résultats

Sommaire

7.1 Introduction

Ce dernier chapitre présente les expérimentations réalisées durant cette thèse afin d'évaluer l'apport scientifique et technique de nos contributions. Nous commençons par une évaluation du système d'extraction d'événements conçu et implémenté tel que décrit dans le chapitre 5. Est ensuite détaillée notre seconde expérimentation appliquée cette fois-ci à l'évaluation du processus d'agrégation sémantique des événements (voir le chapitre 6). Pour chacune des évaluations, nous décrivons, tout d'abord, le système évalué et les différents corpus et paramètres utilisés pour l'expérimentation. Puis, nous présentons une analyse qualitative et/ou quantitative (par différentes métriques) des résultats obtenus. Enfin, nous exposons les différentes limites et perspectives de ces évaluations.

7.2 Évaluation du système d'extraction

La première expérimentation vise à évaluer l'approche d'extraction des événements élaborée et détaillée en section 5.4. Nos objectifs sont, d'une part, d'estimer de façon quantitative les performances de chacune des méthodes d'extraction d'événements conçue et implémentée (l'extracteur à base de règles et celui fondé sur un apprentissage de motifs) et, d'autre part, de montrer leur complémentarité et l'utilité de les combiner.

7.2.1 Protocole d'évaluation

Pour réaliser ces objectifs nous proposons d'extraire automatiquement l'ensemble des événements d'intérêt pour notre domaine d'application au sein d'une collection de textes de type journalistique en anglais et dont le thème est d'intérêt pour le ROSO. Nous nous focalisons donc sur la vingtaine de types d'événement listée en section 4.2.1 et sur leurs dimensions : temporelle (la date de l'événement), spatiale (son lieu d'occurrence) et agentive (les personnes et organisations impliquées) conformément au modèle d'événement présenté au chapitre 4.2. Nous présentons par la suite les données et les paramètres d'apprentissage ayant servi à mettre en place l'approche à base de motifs séquentiels. Précisons que la première approche symbolique n'a pas nécessité de paramétrage particulier.

Données d'apprentissage

Le premier ensemble de données nécessaire est un corpus d'apprentissage afin de mettre en œuvre notre système d'extraction par motifs séquentiels fréquents. Conformément aux paramètres définis en section 5.4.2, la découverte de motifs pertinents pour notre tâche d'extraction des événements nécessite un corpus de textes présentant les caractéristiques suivantes :

- découpage en phrases ;
- découpage en *tokens* ;
- lemmatisation ;
- étiquetage grammatical ;
- annotation des entités nommées (dates, lieux, personnes et organisations) ;
- repérage des déclencheurs d'événements.

Nous avons constitué ce corpus de manière semi-automatique à partir de 400 dépêches de presse sur l'engagement du Canada en Afghanistan collectées sur le Web [116] et de 700 dépêches parues entre 2003 et 2009 sur le site de l'ISAF [117]. Dans un second temps, nous avons traité automatiquement ce corpus pour obtenir l'ensemble des annotations nécessaires : les découpages en phrases et mots ainsi que la lemmatisation ont été réalisés par des modules fournis dans GATE, l'étiquetage grammatical par l'outil TreeTagger et enfin l'annotation en entités nommées et déclencheurs d'événements grâce à notre système d'extraction d'EN et nos *gazetteers* d'événements. Puis, nous avons révisé manuellement ces deux derniers types d'annotation afin de corriger les éventuelles erreurs et ainsi garantir une meilleure qualité des données d'apprentissage. Enfin, un ensemble de pré-traitements spécifiques sont réalisés pour transformer l'ensemble de ce corpus au format supporté par l'outil d'extraction de motifs (notamment la constitution des *itemsets*). L'annexe K présente, à titre d'exemples, plusieurs phrases extraites du corpus d'apprentissage.

Données de test

Le second jeu de données nécessaire est un corpus de test (ou référence) 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. Il est fourni avec ce corpus un ensemble de fiches de référence, une seule fiche correspondant à une seule dépêche et décrivant l'événement principal relaté par celle-ci ainsi qu'un ensemble de propriétés. Ces fiches d'événements ne correspondant pas exactement aux besoins de cette évaluation, nous n'avons pu les réutiliser comme référence. En effet, la granularité (une seule fiche d'événement est proposée pour la totalité d'une dépêche) et le type des événements (uniquement de type *ATTACK* et *BOMBING*) ne correspondent pas à notre modélisation et ne permettent pas d'évaluer la totalité de notre approche. Par conséquent, notre évaluation porte sur une partie de ce corpus qui a été annotée manuellement, soit environ 210 mentions d'événements et près de 240 de leurs dimensions (55 dimensions temporelles, 65 dimensions spatiales et 120 dimensions agentives). L'annotation manuelle a été facilitée notamment par l'utilisation de la plateforme Glozz [118] dédiée à l'annotation et à l'exploration de corpus. L'annexe L présente, à titre d'exemple, une dépêche du corpus de test annotée avec les événements de référence.

Phase d'apprentissage

Nous avons, tout d'abord, opéré un apprentissage de motifs séquentiels fréquents sur le premier corpus en considérant quatre types d'item : la forme fléchie du mot, sa catégorie grammaticale, son lemme et sa classe sémantique (*LookupEvent*, *Date*, *Place*, Unit ou *Person*). Nous avons choisi de réaliser une tâche d'apprentissage par type de d'entité impliquée en utilisant le système des contraintes d'appartenance. Nous avons donc défini et employées les quatre contraintes :

– Le motif retourné doit contenir un déclencheur d'événement et au minimum une entité de type *Date* ;

116. http://www.afghanistan.gc.ca/canada-afghanistan, consulté le 21/03/2012
117. International Security Assistance Force, http://www.nato.int/isaf/docu/pressreleases, consulté le 21/03/2012
118. http://www.glozz.org/

– Le motif retourné doit contenir un déclencheur d'événement et au minimum une entité de type *Place* ;
– Le motif retourné doit contenir un déclencheur d'événement et au minimum une entité de type *Unit* ;
– Le motif retourné doit contenir un déclencheur d'événement et au minimum une entité de type *Person*.

Nous obtiendrons donc après apprentissage quatre ensembles de motifs de type LookupEvent-Date, LookupEvent-Place, LookupEvent-Person et LookupEvent-Unit.

Au préalable, il est nécessaire de fixer les deux paramètres de support et de *gap* (voir la section 5.4.2) : le premier donnant le nombre d'occurrences minimal du motif dans l'ensemble des séquences et le second permettant de "relâcher" la composition des motifs en autorisant un nombre maximal de mots possibles entre chaque élément du motif. Ce paramétrage a été effectué de façon itérative : nous avons fixé une première de valeur de support et de *gap* pour chaque apprentissage, effectué une première extraction de motifs, estimé leur qualité (en observant leur nombre, leur généricité/spécificité, leur couverture/précision), puis ajusté ces paramètres en fonction de nos observations. Le tableau 7.2 présente le nombre de motifs retournés par type de motif en fonction des paramètres choisis. Au regard de ces tests, nous avons choisi de fixer un *gap* maximal de 3 (correspondant à 3 mots possibles entre chaque élément du motif) et un support absolu relativement bas (environ 6% des séquences) afin d'obtenir des motifs intéressants mais en nombre raisonnable pour une exploration et une validation manuelles (environ 12000 motifs au total). Une fois l'ensemble des motifs raffinés par les contraintes d'appartenance et de *gap*, nous avons sélectionnés manuellement et les plus pertinents grâce à l'outil Camelis. La figure 7.1 suivante présente certains des motifs retenus [119], nous en avons conservés au total une cinquantaine.

```
1  { Person }{ VBD }{ DT }{ NN }{ IN }{ DT }{ Event }
2  { , }{ DT }{ NN }{ Event }{ JJ }{ IN }{ Place }
3  { JJ }{ Event }{ NP }{ NP }{ Date }
4  { serve VBG }{ IN as }{ a }{ NN }{ IN }{ the }{ JJ }{ NP Unit }{ Event }
```

FIGURE 7.1 – Exemples de motifs séquentiels fréquents sélectionnés

Pour finir, le jeu de motifs final a été converti en un ensemble de règles JAPE, nous obtenons donc un module indépendant facilement intégrable au sein de notre chaine globale d'extraction d'événements.

Évaluation réalisée

Une fois cette phase d'apprentissage accomplie, nous avons constitué trois chaines d'extraction GATE distinctes :

Chaine 1 Approche à base de règles linguistiques (voir la section 5.4.1) ;

Chaine 2 Approche par apprentissage de motifs (voir la section 5.4.2) ;

119. Les items des motifs correspondent aux catégories grammaticales et aux annotations sémantiques telles que *VBD* : auxiliaire "être" au passé, *DT* : déterminant, *NN* : nom commun au singulier, *IN* : préposition, *JJ* : adjectif, *NP* : nom propre au singulier, *VBG* : auxiliaire "être" au participe présent, *Person* : entité nommée de type personne, *Event* : déclencheur d'événement équivalant à l'étiquette *LookupEvent* ci-dessus, *Place* : entité nommée de type lieu, *Date* : entité nommée de type date, *Unit* : entité nommée de type organisation.

	LookupEvent-Date	*LookupEvent-Place*	*LookupEvent-Person*	*LookupEvent-Unit*
Sup3 Gap0	-	-	-	1381
Sup4 Gap0	113	-	67748	-
Sup4 Gap2	-	-	-	18278
Sup6 Gap0	-	1000	-	-
Sup6 Gap2	1046	-	-	2039
Sup8 Gap0	30	317	-	-
Sup8 Gap2	699	-	-	725
Sup8 Gap3	1108	-	-	-
Sup10 Gap2	-	8693	1730	344
Sup10 Gap3	681	6596	9540	47
Sup12 Gap3	-	-	4614	-

FIGURE 7.2 – Nombre de motifs retournés en fonction des paramètres choisis

Chaine 3 Union des deux approches : elle contient l'ensemble des pré-traitements nécessaires à chaque approche (listés en section 5.4) ainsi que les deux modules de règles développés exécutés successivement.

Ces trois chaines ont été exécutées sur le corpus de test et nous avons comparé manuellement et par document la qualité des événements extraits aux événements de référence correspondants. Cela nous a permis de calculer des scores de précision, rappel et F1-mesure (voir la section 2.5.1 pour la définition de ces métriques) pour chaque chaine.

Par ailleurs, nous avons souhaité évaluer l'influence de l'analyse syntaxique en dépendance et de la qualité de la REN sur nos systèmes d'extraction d'événements. Le tableau 7.1 ci-dessous résume les différentes variantes des chaines évaluées et dont les résultats sont présentés dans la section suivante.

	Règles manuelles	Motifs séquentiels	Union des résultats
Avec analyse syntaxique	x		x
Sans analyse syntaxique	x		x
REN automatique	x	x	x
REN manuelle	x	x	x

TABLE 7.1 – Chaines d'extraction d'événements : variantes évaluées

7.2.2 Analyse des résultats

Le tableau 7.2 présente les scores de précision, rappel et F1-mesure obtenus : il s'agit des résultats d'évaluation des 3 chaines présentées plus haut par type de dimension et toutes dimensions confondues. Précisons qu'il s'agit des métriques calculées avec une annotation manuelle des entités nommées pour les 3 chaines et avec analyse syntaxique en dépendance pour la chaine 1.

Nous pouvons tout d'abord constater que l'approche à base de règles et l'apprentissage de motifs obtiennent tous deux une très bonne précision globale et que, 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

	Approche à base de règles manuelles			Apprentissage de motifs			Union des résultats		
	Précision	*Rappel*	*F1-mesure*	*Précision*	*Rappel*	*F1-mesure*	*Précision*	*Rappel*	*F1-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 dimensions	0,94	0,37	**0,45**	0,90	0,48	**0,62**	0,88	0,60	**0,71**

TABLE 7.2 – Extraction d'événements : précision, rappel et F-mesure

apprentissage, que nous expliquons par une sélection manuelle restrictive et précise des motifs. Quant aux taux de rappel peu élevés, ce n'est pas rare pour les approches à base de règles construites manuellement et, dans le cas de l'apprentissage de motifs, cela peut être du à un "gap" maximal trop restreint qui ne permet pas d'extraire les relations distantes. Par ailleurs, une analyse des résultats par type de dimension permet d'en apprendre sur les points forts et faiblesses de chacune des approches développées. Nous pouvons, par exemple, voir que le système à base de motifs séquentiels est nettement plus performant pour la reconnaissance des dates des événements. Concernant la dimension spatiale, bien que celle-ci soit meilleure en termes de F1-mesure, l'approche symbolique s'avère plus précise. Enfin, tandis que pour ces deux premières dimensions (temporelle et spatiale) l'union des résultats apporte peu par rapport à la meilleure des deux approches, celle-ci s'avère particulièrement utile pour la dimension agentive.

Enfin, nous pouvons retenir de cette expérimentation que l'union des résultats obtient une F1-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 dimension. 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 tout de même une légère perte de précision de l'union par rapport aux deux approches seules : celle-ci résulte du fait que réaliser une simple union des résultats, même si elle permet d'additionner les vrais positifs des deux approches, entraine aussi une addition des faux positifs. Toutes approches confondues, les résultats obtenus sont satisfaisants comparés à l'état de l'art même s'il convient de prendre des précautions lorsque l'on compare des systèmes évalués selon différents protocoles.

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 symbolique. Le tableau 7.3 présente les performances de notre système avec et sans analyse syntaxique en dépendance. Nous pouvons constater que celle-ci permet d'augmenter sensiblement la qualité des extractions pour tous les types de dimension. Bien que les outils d'analyse syntaxique soient inégalement disponibles selon les langues, cette observation confirme l'intérêt de cette technique pour l'extraction des événements. Une perspective intéressante serait d'exploiter les résultats de l'analyse syntaxique au sein de la seconde approche en tant que caractéristique supplémentaire à prendre en compte lors de l'apprentissage des motifs séquentiels fréquents.

Influence de la reconnaissance automatique des entités nommées

Pour compléter les résultats précédents fondés sur une annotation manuelle des entités nommées, nous avons évalué les 3 chaines avec une annotation automatique des entités (voir la section 5.3). En

	Sans analyse syntaxique			Avec analyse syntaxique		
	Précision	*Rappel*	*F1-mesure*	*Précision*	*Rappel*	*F1-mesure*
Date	0,32	0,45	0,38	0,93	0,25	0,39
Lieu	0,86	0,18	0,30	0,92	0,37	0,53
Participants	0,90	0,31	0,46	0,97	0,49	0,42
Toutes dimensions	0,69	0,31	**0,38**	0,94	0,37	**0,45**

TABLE 7.3 – Extraction d'événements : apport de l'analyse syntaxique

effet, les entités nommées étant repérées de façon automatique dans les applications réelles, il est important d'estimer quelle sera l'influence de cette automatisation sur l'extraction des événements. Le tableau 7.4 ci-dessous compare les performances de la 3ème chaine d'extraction (c'est-à-dire l'union des deux approches) avec une REN réalisée manuellement ou automatiquement. Nous observons une baisse générale de la qualité des extractions qui s'explique par le fait que notre système d'extraction des événements dépend de ces entités nommées pour la reconnaissances des différentes dimensions. Par conséquent, si une entité nommée impliquée dans un événement n'a pas été reconnue, celui-ci ne pourra la reconnaître en tant que dimension de cet événement (ce qui fait diminuer le rappel de notre approche). Par ailleurs, si une entité a bien été reconnue mais mal catégorisée ou mal délimitée cela entrainera une perte de précision pour notre extracteur d'événements. Malgré cette influence négative sur les résultats globaux, les scores obtenus par notre approche restent à la hauteur de l'état de l'art.

	REN manuelle			REN automatique		
	Précision	*Rappel*	*F1-mesure*	*Précision*	*Rappel*	*F1-mesure*
Date	0,90	0,68	0,78	0,86	0,64	0,73
Lieu	0,81	0,60	0,69	0,94	0,46	0,62
Participants	0,92	0,51	0,66	0,94	0,39	0,55
Toutes dimensions	0,88	0,60	**0,71**	0,91	0,50	**0,64**

TABLE 7.4 – Extraction d'événements : influence de la REN

7.2.3 Bilan de l'évaluation

Cette première évaluation a permis les observations suivantes :

1. les deux approches proposées pour l'extraction des événements obtiennent des résultats équivalents voire supérieurs aux systèmes existants (voir la section 2.5.2) ;
2. l'analyse syntaxique en dépendance des phrases améliore significativement la qualité des extractions réalisées par l'approche symbolique ;
3. les performances globales de notre approche sont impactées à un niveau acceptable par la détection automatique des entités nommées ;
4. les deux techniques mises en œuvre pour l'extraction des événements présentent des forces complémentaires ;
5. une union de leurs résultats (peu coûteuse à réaliser) permet d'améliorer sensiblement la qualité des événements extraits.

Nous dressons donc un bilan positif de cette première évaluation. Toutefois, celle-ci présente des limites qui donnent lieu à plusieurs perspectives d'amélioration. Tout d'abord, la conclusion 1 bien que vraie si l'on compare les scores obtenus, doit être nuancée car il s'avère toujours difficile de comparer des systèmes d'extraction ayant été évalués selon des protocoles différents. L'influence de la REN automatique révèle une problématique plus largement présente dans le contexte du *Media Mining* : il s'agit de l'interdépendance des modules de traitement et du suivi de qualité tout au long de la chaine de traitement. Dans notre cas de figure, il serait intéressant de mettre en place une réelle collaboration (et non plus une simple juxtaposition) entre les modules de REN et d'extraction d'événements en explorant, par exemple, les travaux sur la qualité des annotations. Par ailleurs, suite à l'observation 2, nous souhaiterions améliorer notre approche d'extraction de motifs séquentiels avec de nouvelles caractéristiques issues d'une analyse syntaxique. Celles-ci pourraient notamment servir en tant que contraintes supplémentaires afin de limiter la quantité de motifs retournés. Enfin, évaluer les résultats obtenus par une simple union des deux approches a mis en exergue leurs complémentarités et de nouvelles pistes d'exploration pour élaborer une combinaison plus adaptée. Dans le cadre de nos travaux, nous avons choisi de réaliser l'extraction selon les deux approches successivement et de gérer cette problématique sur le même procédé que les mentions d'événements provenant de différents documents.

7.3 Premières expérimentions sur l'agrégation sémantique

Cette seconde expérimentation vise à évaluer le prototype d'agrégation sémantique des événements que nous avons développé selon l'approche présentée au chapitre 6. Nous présentons, dans un premier temps, les principales caractéristiques techniques de ce prototype puis le jeu de données employé pour cette expérimentation. Par la suite, nous présentons nos premiers résultats et concluons par un bilan de cette expérimentation ainsi que les perspectives envisagées.

7.3.1 Implémentation d'un prototype

L'approche proposée au chapitre 6 a été implémentée au sein de plusieurs services (en langage Java) permettant de traiter un ensemble de documents au format WebLab issus du service d'extraction d'information développé (voir la section 5.4). Un exemple de document contenant des événements extraits par notre système et représentés en RDF/XML selon le schéma l'ontologie WOOKIE est proposé en annexe I. L'ensemble des connaissances créé et modifié par ces services est stocké et géré au sein de bases de connaissances sémantiques grâce au *triplestore* Jena Fuseki [120]. Dans un premier temps, les événements extraits ainsi que leurs dimensions provenant du système d'extraction sont stockés dans une première base de connaissances A régie par notre ontologie de domaine (voir la section 4.3).

Les différents calculs de similarité ont été implémentés au sein d'un premier service qui présente les fonctionnalités suivantes (le reste des fonctions restant comme perspectives à nos travaux) :

- similarité conceptuelle : implémentation par des tests de subsomption ontologique (telle que présentée en section 6.3.1) ;
- similarité temporelle : implémentation telle que proposée en section 6.3.2 mais sans la fonction de distance temporelle ;

120. http://jena.apache.org/documentation/serving_data/

– similarité spatiale : implémentation du service de désambiguïsation spatiale par GeoNames (voir la section 6.3.3) mais sans la fonction de distance spatiale ;
– similarité agentive : implémentation par distance de Jaro-Winkler comme proposé en section 6.3.4, la désambiguïsation avec DBPedia reste à mettre en place.

Les calculs de similarité sont combinés grâce à la librairie Apache Jena [121] et son mécanisme de règles d'inférence (voir l'annexe J pour un exemple de règle). Le moteur d'inférence développé est appliqué à la base A qui se trouve ainsi augmentée de l'ensemble des liens de similarité.

Un second service réalise le processus d'agrégation sémantique : celui-ci permet de définir une configuration et d'appliquer une phase de regroupement au graphe de similarité entre événements (chargé dans la base A). Une fois le regroupement réalisé, c'est-à-dire lorsque le premier graphe a été enrichi avec les agrégats d'événements similaires, celui-ci est stocké dans une seconde base de connaissances B. Cette base sera alors disponible et interrogeable par les services de la plateforme WebLab pour présenter les agrégats à l'utilisateur final par divers modes de visualisation.

Cette implémentation constitue une première preuve de concept pour montrer la faisabilité de notre processus d'agrégation sémantique. Toutefois, le système que nous avons conçu n'est, à l'heure actuelle, pas apte à passer à l'échelle pour obtenir des résultats significatifs sur un corpus d'évaluation à taille réelle. Ce passage à l'échelle constitue la principale perspective à court terme de nos travaux. Nous présentons dans les sections suivantes les observations que nous avons pu réaliser sur un jeu de données réelles plus réduit.

7.3.2 Jeu de données

Pour cette expérimentation, nous nous appuyons sur une base de données nommée Global Terrorism Database (GTD). Il s'agit d'une base open-source contenant plus de 104 000 événements terroristes recensés manuellement de 1970 à 2011 [LaFree and Dugan, 2007]. Cette collection est gérée par le consortium américain START [122] étudiant les faits terroristes survenus dans le monde. Celle-ci est constituée à la fois d'événements d'échelle internationale et nationale, principalement collectés à partir de sources d'actualité sur le Web mais aussi provenant de bases de données existantes, livres, journaux et documents légaux. Les événements dans la base GTD sont catégorisés selon les neuf types suivants :

1. Assassination
2. Hijacking
3. Kidnapping
4. Barricade Incident
5. Bombing/Explosion
6. Unknown
7. Armed Assault
8. Unarmed Assault
9. Facility/Infrastructure Attack

121. https://jena.apache.org/
122. Study of Terrorism and Responses to Terrorism

En fonction de son type, un événement peut présenter entre 45 et 120 propriétés/attributs tels que les suivants :

- type de l'événement ;
- date de l'événement (année, mois, jour) ;
- lieu de l'événement (région, pays, état/province, ville, latitude/longitude) ;
- auteur(s) (personne ou groupe) ;
- nature de la cible ;
- type de l'arme utilisée ;
- dommages matériels ;
- nombre de décès ;
- nombre de blessés ;
- résumé ;
- indice de certitude ;
- sources (1 à 3) : nom de la source, extrait (titre), date de l'article, URL ;
- etc.

L'ensemble de ces données est représenté sous forme de couples "champ-valeur" et téléchargeable au format CSV. Nous disposons d'ores et déjà d'une partie de cette base (environ 4800 fiches d'événements survenus en 2010) convertie en base de connaissance sémantique (graphe RDF).

Cette base de données est bien adaptée à l'évaluation de notre processus global de capitalisation des connaissances car elle constitue une collection de fiches d'événements manuellement agrégées à partir de plusieurs sources d'information. Les sources (articles et dépêches principalement) à l'origine d'une fiche sont accessibles via l'attribut *scite* qui spécifie leurs URLs. Une collecte automatique de ces sources (grâce à un service WebLab) permet donc facilement de constituer un corpus d'évaluation composé de dépêches de presse et des fiches d'événements correspondantes.

Il faut noter également que cette base est fondée sur une modélisation différente de celle définie dans le cadre de nos recherches (l'ontologie WOOKIE). Il nous faut donc trouver le meilleur alignement de classes et attributs afin de pouvoir évaluer les résultats de notre approche par rapport aux fiches de référence de la base GTD. Le modèle d'événement employé par cette base étant sensiblement similaire au nôtre, l'alignement des attributs d'intérêt (date, lieu et participants) n'a pas soulevé de difficultés. Concernant la taxonomie des événements, 4 classes (sur les 9 du modèle GTD) ont pu être alignées avec le modèle WOOKIE car ayant la même sémantique (voir le tableau 7.5). Nos expérimentations sont donc limitées par ce point et ne concernent que des événements de ces 4 types.

Modèle GTD	**Ontologie WOOKIE**
Facility/Infrastructure Attack	DamageEvent
Bombing/Explosion	BombingEvent
Armed Assault	AttackEvent
Hostage Taking	KidnappingEvent

TABLE 7.5 – Alignement des types d'événement entre le modèle GTD et l'ontologie WOOKIE

7.3.3 Exemples d'observations

Calculs de similarité

Nous présentons ici un exemple d'application à ce jeu de données de la similarité sémantique entre événements : nous collectons et analysons trois dépêches de presse (que nous nommerons *source1*, *source2* et *source3*) à l'origine d'une même fiche d'événement issue de la base GTD [123] et résumée par la figure 7.3. Le tableau suivant 7.6 présente quatre des événements extraits automatiquement par

eventid	201110310005
iyear	2011
imonth	10
iday	31
country_txt	Kazakhstan
region_txt	Central Asia
city	Atyrau
attacktype1_txt	Bombing/Explosion
target1	A civilian residence was targeted.
gname	Soldiers of the Caliphate
scite1	Russia-Eurasia Terror Watch, "Killing Spree Leaves Seven Dead in Taras," Russia-Eurasia Terror Watch, November 12, 2011, http://www.retwa.com/home.cfm?articleId=12303.
scite2	Radio Free Europe, "Kazakh Police Apprehend Three Over Bomb Blasts," Radio Free Europe, November 7, 2011, http://www.rferl.org/content/kazakh_police_apprehend_three_over_bomb_blasts/24383574.html.
scite3	Trend News Agency, ""Soldiers of Caliphate" Claims Responsibility for Blasts in Kazakhstan," Trend News Agency, November 1, 2011, http://en.trend.az/regions/casia/kazakhstan/1952405.html.

FIGURE 7.3 – Un exemple d'événement issu de la base GTD

notre système à partir de ces sources, ainsi que leurs dimensions.

	Event1	**Event2**	**Event3**	**Event4**
rdfs :label	explosions	explosions	bomb blasts	attacked
rdf :type	BombingEvent	BombingEvent	BombingEvent	AttackEvent
wookie :date	-	[∅ ,10,31]	-	-
wookie :takesPlaceAt	Atirau city	-	-	Kazakstan
wookie :involves	-	KNB department	Kazakh Police	-
source	scite1	scite2	scite3	scite2

TABLE 7.6 – Événements extraits et leurs dimensions

Nous constatons dans cet exemple que, sans post-traitement adapté, ces quatre événements seraient présentés à l'utilisateur final de façon distincte et sans aucun lien entre eux (quatre fiches de connaissance différentes seraient créées). Toutefois, appliquer notre modèle de similarité sémantique entre événements, permet de compléter cet ensemble de connaissances avec des relations de similarité entre ces quatre fiches de connaissance. La figure 7.4 constitue un extrait des différentes similarités obtenues représentées en RDF/XML. Tout d'abord, une proximité conceptuelle a été détectée entre les événements *Event2* et *Event4* (ligne 9) de par le fait que le type de l'événement *Event2* est une sous-classe de celui de l'événement *Event4* dans l'ontologie de domaine. De plus, le service de désambiguïsation géographique ayant assigné des identifiants GeoNames aux entités *Atirau city* et *Kazahstan*, nous pouvons appliquer

123. National Consortium for the Study of Terrorism and Responses to Terrorism (START). (2012). Global Terrorism Database [Data file]. Retrieved from http://www.start.umd.edu/gtd

le calcul de similarité spatiale et déduire que les événements *Event1* et *Event4* sont spatialement proches (ligne 5). Enfin, la fonction de similarité agentive utilisant la distance de Jaro-Winkler a permis d'estimer que les participants des événements *Event2* et *Event3* ne co-réfèrent pas (ligne 10).

```
1  <rdf:Description rdf:about="#Event1">
2    <wookie:isConceptuallyID rdf:resource="#Event2"/>
3    <wookie:isConceptuallyID rdf:resource="#Event3"/>
4    <wookie:isConceptuallyPROX rdf:resource="#Event4"/>
5    <wookie:isSpatiallyPROX rdf:resource="#Event4"/>
6  </rdf:Description>
7  <rdf:Description rdf:about="#Event2">
8    <wookie:isConceptuallyID rdf:resource="#Event3"/>
9    <wookie:isConceptuallyPROX rdf:resource="#Event4"/>
10   <wookie:isAgentivelyDIFF rdf:resource="#Event3"/>
11 </rdf:Description>
12 <rdf:Description rdf:about="#Event3">
13   <wookie:isConceptuallyPROX rdf:resource="#Event4"/>
14 </rdf:Description>
```

FIGURE 7.4 – Similarités entre événements : extrait représenté en RDF/XML

Processus d'agrégation

Dans un second temps, notre processus global d'agrégation a été testé sur un sous-ensemble du corpus GTD : l'ensemble des traitements (extraction, calculs de similarité et agrégation sémantique) a été appliqué sur 60 dépêches de presse collectées du Web. Nous pouvons déjà faire quelques observations chiffrées sur ce corpus de test et les résultats obtenus :

(a) 92 fiches d'événements de référence correspondent à ces 60 dépêches ;

(b) la majorité des dépêches (40) est associée à une seule fiche de référence, 15 dépêches relatent chacune 2 événements de référence et les 5 restantes renvoient chacune à 3 événements ;

(c) 223 mentions d'événements ont été repérées par notre approche d'extraction (tous types confondus) ;

(d) 44 de ces événements extraits font partie des 4 types de l'alignement (voir le tableau 7.5 et ont pu donc être évalués ;

(e) ces 44 mentions d'événements comprennent 20 *AttackEvent*, 14 *BombingEvent*, 2 *DamageEvent* et 8 *KidnappingEvent* ;

(f) parmi ces 44 mentions, 3 possèdent une date, 7 présentent un lieu extrait et 5 impliquent des participants.

Nous nous sommes ensuite intéressés aux différentes relations de similarité de type *ID* et *PROX* (les plus pertinentes pour rapprocher des événements similaires) créées entre les 44 événements extraits :

- 20 relations de type *isConceptuallyID* ont été détectées ;
- 17 relations de type *isConceptuallyPROX* ;
- 4 relations de type *isAgentivelyID* ;
- 3 relations de type *isSpatiallyPROX*.

Nous pouvons constater que peu de relations de ce type ont été détectées. Cela est, tout d'abord, à corréler avec l'observation (f) faite plus haut, montrant que, parmi les 44 événements extraits et analysables, peu de dimensions ont été extraites. Ce manque est du notamment à une limite de notre corpus d'évaluation mise en avant par l'observation (b) : chaque événement du jeu de données est rapporté au maximum par 3 sources, ce qui ne reflète pas les conditions réelles d'un processus de veille, où de nombreuses sources et articles reportant le même événement sont quotidiennement collectés et analysés.

Dans un second temps, nous avons appliqué un premier regroupement avec pour condition de regrouper entre eux les événements partageant au minimum une similarité de niveau *ID* et une similarité de niveau *PROX*. Au vu des relations de similarité ci-dessus, seules deux configurations ont permis d'obtenir des agrégats :

- $\{isConceptuallyID, isTemporallyLI, isSpatiallyLI, isAgentivelyID\}$ produit 3 agrégats d'événements ;
- $\{isConceptuallyPROX, isTemporallyLI, isSpatiallyPROX, isAgentivelyLI\}$ produit 1 agrégat d'événements.

Appliqué sur les 3 agrégats produits par la première configuration, un second regroupement par la configuration $\{isConceptuallyDIFF, isTemporallyLI, isSpatiallyLI, isAgentivelyID\}$ permet d'obtenir un agrégat contenant les trois événements présentés dans le tableau 7.7. Deux des événements proviennent d'une même source (*Event2* et *Event3*) et le troisième (*Event1*) d'une source différente, celles-ci sont reportées en annexes M et N.

	Event1	**Event2**	**Event3**
gtd :source	s3	s12	s12
rdfs :label	kidnapping	kidnapped	bomb
wookie :involves	Taliban		Taliban

TABLE 7.7 – Exemple de 3 événements agrégés automatiquement

Afin d'évaluer l'apport de cette agrégation, nous avons examiné les fiches de référence associées aux deux sources dont ces événements ont été extraits : la source *s3* renvoie à un seul événement (que nous nommerons *Reference1*) et la source *s12* renvoie à deux événements (que nous nommerons *Reference2* et *Reference3*). Le tableau 7.8 ci-dessous présente ces trois fiches de référence ainsi que quelques propriétés d'intérêt pour cette expérimentation.

Nous pouvons, à partir de cet exemple, faire les observations suivantes :

- Les trois extractions d'événements réalisées correspondent bien aux fiches de référence associées à leurs sources respectives ;
- Les deux regroupements successifs ont permis d'agréger 3 événements perpétrés par le même agent ("Taliban") ;
- Une géolocalisation des 3 événements de référence (voir la figure 7.5) montrent que les 3 événements agrégés se sont produits dans la même zone géographique. Bien que les lieux des 3 événements n'aient pas été repérés automatiquement, l'agrégation permettrait à l'analyste de découvrir cette proximité (en remontant par exemple aux sources des trois événements) ;
- Dans l'hypothèse qu'avec une plus grande quantité de sources analysées les dimensions manquantes (date et lieu) auraient pu être extraites, l'analyste aurait pu analyser la répartition spatiale et temporelle de ces 3 événements afin d'en déduire de nouvelles connaissances ou de nouvelles prédictions.

	Reference1	Reference2	Reference3
eventID	201009260002	201001230002	201001230007
source	s3	s12	s12
year	2010	2010	2010
month	9	1	1
day	26	23	23
country	Afghanistan	Afghanistan	Afghanistan
region	South Asia	South Asia	South Asia
province/state	Konar	Konar	Paktika
city	Chawkay	Shigal	Unknown
attackType1	Hostage Taking (Kidnapping)	Hostage Taking (Kidnapping)	Armed Assault
attackType2	Armed Assault		
targetType1	NGO	Police	Transportation
corp1	Development Alternatives Inc	Sheigal Law Enforcement	
target1	Linda Norgrove	police chief of Sheigal district and two other police officers	Civilians
targetType2	NGO		
corp2	Development Alternatives Inc		
target2	Three Afghan aid employees		
perpetrator	Taliban	Taliban	Taliban

TABLE 7.8 – Fiches d'événements de référence

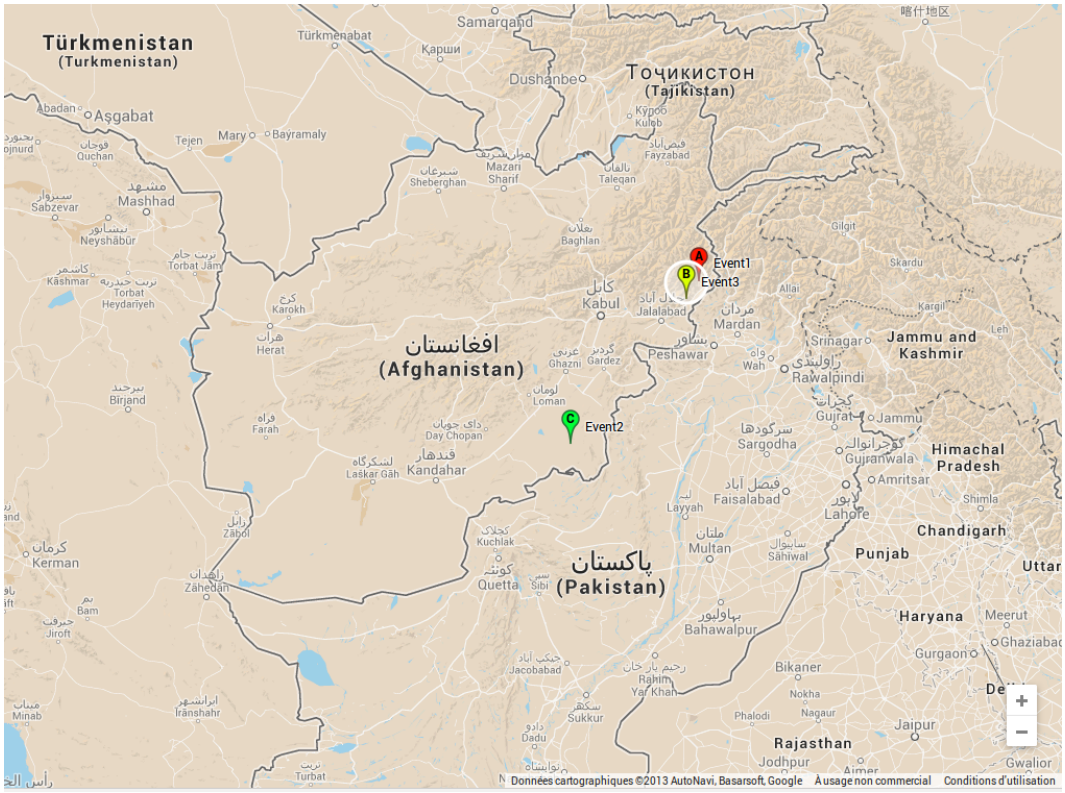


FIGURE 7.5 – Visualisation des 3 événements extraits sur une carte géographique

7.3.4 Bilan de l'expérimentation

Cette première expérimentation a montré des résultats prometteurs à la fois pour notre approche de similarité entre événements et pour le processus d'agrégation sémantique en tenant compte de ces similarités. En effet, nous avons montré que notre approche globale peut être mise en œuvre sur des

données réelles sans nécessiter d'effort d'adaptation important et en maintenant le niveau de qualité escompté.

Toutefois, celle-ci présente un certain nombre de limites qui constituent autant de perspectives à explorer à l'avenir. Tout d'abord, nous souhaitons améliorer notre prototype en implémentant la totalité des mesures de similarité proposées en section 6.3 (distances temporelles et spatiales ainsi que la desambiguïsation agentive grâce à la base DBPedia). De plus, nous complèterons le service de regroupement en y intégrant le procédé de regroupement hiérarchique par plusieurs passes (une seule passe est réalisée pour le moment). Enfin, des problèmes techniques empêchent, à l'heure actuelle, d'obtenir des résultats significatifs et quantitatifs sur un plus grand ensemble de données. Ce passage à l'échelle constitue notre plus proche perspective future et permettra d'évaluer notre approche de façon exhaustive et avec des métriques adaptées (issues par exemple de l'évaluation des méthodes de *clustering*) sur l'ensemble du jeu de données présenté en section 7.3.2 : soit environ 5000 événements de référence. Cette évaluation devra principalement permettre de répondre aux deux questions suivantes :

- est-ce que les événements extraits correspondant à une seule et même fiche de référence sont bien rapprochés ?
- est-ce que les événements extraits ne correspondant pas à la même fiche de référence sont bien différenciés ?

7.4 Conclusions

Dans ce dernier chapitre, nous avons présenté deux expérimentations destinées à évaluer deux de nos contributions. Tout d'abord, une première évaluation a porté sur notre approche mixte d'extraction automatique des événements : celle-ci a été testée et comparée à un corpus de référence annoté manuellement pour les besoins de l'expérimentation. Pour ce faire, nous avons effectué une évaluation en termes de précision/rappel/F1-mesure et cela selon différentes configurations : chaque approche a été évaluée séparément puis comparée à l'union de leurs résultats. Nous avons également fait varier différents paramètres tels que la présence de l'analyse syntaxique ou encore l'automatisation de la REN. Une analyse détaillée des résultats a montré de bonnes performances en comparaison avec l'état de l'art et ceci pour l'ensemble des configurations d'évaluation testées. Cette première évaluation a également pointé quelques limites de notre approche telles que l'impact de l'extraction d'entités nommées sur les performances des extracteurs d'événements. De plus, celle-ci pourrait être améliorée en exploitant davantage les informations fournies par l'analyse syntaxique en dépendance.

La seconde expérimentation a constitué une première analyse qualitative dans le but d'illustrer les résultats de notre processus d'agrégation sémantique sur un jeu de données réelles. Nous avons présenté, dans un premier temps, l'implémentation d'un prototype fonctionnel couvrant la chaine complète d'agrégation des événements. Puis, la base de données *Global Terrorism Database* a été introduite ainsi qu'un sous-ensemble de celle-ci servant de corpus d'évaluation pour cette expérimentation. Les tests réalisés se sont avérés prometteurs à la fois pour ce qui est du calcul de similarité entre événements et du processus d'agrégation proposé. Le premier traitement a rapproché efficacement des mentions d'événements qui co-réfèrent et permettra ainsi de réduire la tâche de l'analyste du ROSO en lui proposant l'ensemble des liens de similarité en tant que critères supplémentaires de recherche. Puis, nous avons appliqué notre processus d'agrégation par regroupements successifs sur ce jeu de test et ceci pour pour 3 types de configuration. Malgré le peu d'agrégats formés (en raison de la taille réduite du corpus), nous avons montré

par un exemple l'utilité de cette agrégation du point de vue utilisateur. Cette seconde contribution pourra être améliorée en intégrant l'ensemble des fonctions de similarité proposées à notre prototype d'agrégation et en optimisant celui-ci pour permettre son passage à l'échelle d'un processus de veille réel.

Conclusion et perspectives

Sommaire

1 Synthèse des contributions

La problématique étudiée durant cette thèse est la capitalisation des connaissances à partir de sources ouvertes. Nous nous sommes plus particulièrement intéressés à l'extraction et à l'agrégation automatique des événements dans le domaine du Renseignement d'Origine Sources Ouvertes. Ce sujet de recherche nous a amenés à explorer les principaux axes de recherche suivants :

- La représentation et la modélisation des connaissances ;
- L'extraction automatique d'information ;
- La capitalisation des connaissances.

Pour répondre à cette problématique, nous avons, dans une première phase de nos recherches, réalisé un état de l'art approfondi de ces trois axes scientifiques.

1.1 État de l'art

Nous avons, tout d'abord, exploré les principes théoriques et les recherches actuelles dans les domaines de la représentation des connaissances, du Web sémantique et de la modélisation des événements. Ce premier état de l'art a rappelé la distinction fondamentale entre les concepts de donnée, information et connaissance. Il a également confirmé l'importance croissante des technologies du Web sémantique qui sont partie intégrante d'une majorité de travaux actuels en fouille de documents. Les ontologies, plus particulièrement, se positionnement comme un mode de représentation des connaissances en adéquation avec les nouvelles problématiques du traitement de l'information. Combinées aux bases de connaissances sémantiques et moteurs d'inférence, cela constitue un socle de technologies particulièrement adapté à la capitalisation des connaissances à partir de sources ouvertes. Dans un second temps, nous nous sommes focalisés sur la place des événements en représentation des connaissances et avons étudié les différentes théories, modèles et ontologies proposés jusqu'alors. L'événement apparait comme un concept complexe, dont la définition et la modélisation sont encore aujourd'hui des sujets de discussions et débats. Beaucoup de travaux s'accordent sur un objet multi-dimensionnel impliquant une situation spatio-temporelle et un ensemble d'acteurs et facteurs. Nous avons également exploré les spécifications utilisées par les acteurs du renseignement afin d'adapter notre proposition à ce cadre d'application.

La seconde revue de littérature a été centrée sur l'extraction automatique d'information dans les textes. Celle-ci a révélé un domaine de recherche très étudié bien que relativement jeune : nous avons pu recenser un nombre important d'approches, d'applications possibles, de logiciels et plateformes développés ainsi que de campagnes et projets d'évaluation menés jusqu'à nos jours. Les méthodes développées sont historiquement réparties en deux catégories : les symboliques et les statistiques. Les premières, développées manuellement par des experts de la langue, s'avèrent globalement plus précises, tandis que les secondes réalisent un apprentissage sur une grande quantité de données et présentent généralement un fort taux de rappel. Parallèlement à cela, nous avons constaté une certaine complémentarité des approches existantes, non seulement en termes de précision et de rappel mais également du point de vue des types d'entités ciblées, du genre textuel, du domaine d'application, etc. Il apparait, par conséquent, pertinent de combiner les approches existantes afin de tirer partie de leurs atouts respectifs. Pour ce faire, les approches hybrides constituent des alternatives intéressantes car elles s'avèrent faiblement supervisées et plus flexibles que d'autres approches statistiques. Enfin, ce tour d'horizon nous a permis de comparer

différents outils et logiciels pour la mise en œuvre de notre approche ainsi que différents jeux de données potentiellement adaptés à l'évaluation de nos travaux.

Le dernier état de l'art autour de la capitalisation des connaissances a mis en avant une suite logique à nos travaux en extraction automatique d'information : la conception d'une approche globale permettant la transition du texte vers la connaissance proprement dite. Cette problématique a donné lieu à diverses recherches au sein de plusieurs communautés de l'IA, chacune d'elles manipulant sa propre terminologie adaptée à ses propres besoins. Ses divergences de vocabulaire n'empêchent pas d'observer la place importante réservée à la capitalisation des connaissances au sein des recherches actuelles, que ce soit en réconciliation de données, extraction d'information ou Web sémantique. La majorité des approches proposées en réconciliation de données trouve ses origines dans les bases de données. Il s'agit dans ce cadre d'assurer le maintien de cohérence des bases en détectant les entrées et champs dupliqués. Pour cela, beaucoup de travaux sont fondés sur des calculs de similarité : les plus simples opèrent une similarité entre chaines de caractère tandis que les plus avancés exploitent le contexte linguistique et extra-linguistique des données à comparer. Ces dernières se distinguent ensuite par le type et la méthode employée pour obtenir ces caractéristiques de contexte et sur leur mode de combinaison. Nous avons également exploré les techniques de capitalisation au sein du Web sémantique. A l'heure actuelle, cela est réalisé principalement par ce que l'on nomme la *Wikification* : il s'agit de désambuiguïser sémantiquement les mentions textuelles d'intérêt afin de les rattacher à une entité du monde référencée dans une base de connaissances externe (dans le LOD essentiellement). Enfin, la capitalisation des connaissances sur les événements extraits est l'objet d'un intérêt grandissant. Également nommée co-référence entre événements, cette problématique est adressée principalement par des méthodes statistiques supervisées ou non, d'une part, et des approches fondées sur les graphes et les calculs de similarité d'autre part.

Suite à la réalisation de ces états de l'art, nous avons proposé trois contributions relatives aux trois domaines scientifiques explorés. Celles-ci s'articulent au sein de notre processus global de capitalisation des connaissances. Au regard des conclusions de l'état de l'art, l'objectif directeur de nos recherches est de concevoir un système global de reconnaissance et d'agrégation des événements le plus générique possible et intégrant des méthodes et outils de la littérature.

1.2 Un modèle de connaissances pour le ROSO

Notre première contribution est l'élaboration d'un modèle de connaissances qui servira de guide à notre approche de capitalisation des connaissances. Nous avons, tout d'abord, défini une modélisation des événements fondée sur des modèles reconnus en ingénierie des connaissances et en extraction d'information. Un événement est représenté par quatre dimensions : une dimension conceptuelle (correspondant au type de l'événement), une dimension temporelle (la date/ période d'occurrence de l'événement), une dimension spatiale (le lieu d'occurrence de l'événement) et une dimension agentive (dédiée aux différents acteurs impliqués dans l'événement). Pour la définition de ses dimensions, nous avons privilégié la généricité du modèle ainsi que sa reconnaissances par la communauté scientifique concernée (par exemple, le modèle TUS et la représentation spatiale en aires géographiques). Par ailleurs, afin de modéliser l'ensemble des informations d'intérêt pour les analystes du ROSO, une ontologie de domaine a été proposée. Celle-ci comprend en tant que classes de plus haut niveau les cinq entités principales de ce domaine : les organisations, les lieux, les personnes, les équipements et les événements. De plus, notre ontologie intègre de nombreux liens sémantiques vers d'autres modélisations existantes afin de maintenir une interopérabilité au sein du Web sémantique. Cette contribution présente quelques limites et nous envisageons

des perspectives d'amélioration telles que l'intégration d'une cinquième dimension contextuelle afin de représenter des éléments du contexte linguistique et extra-linguistique, mais également une définition des rôles au sein de la dimension agentive.

1.3 Une approche mixte pour l'extraction automatique des événements

Notre seconde contribution est une approche mixte pour l'extraction automatique des événements fondée sur une méthode symbolique à base de grammaires contextuelles, d'une part, et sur une technique de fouille de motifs séquentiels fréquents, d'autre part. La méthode symbolique comporte un ensemble de règles d'extraction développées manuellement et exploite notamment la sortie d'une analyse syntaxique en dépendance combinée avec un ensemble de classes argumentales d'événement. La seconde technique permet d'obtenir de manière faiblement supervisée un ensemble de patrons d'extraction grâce à la fouille de motifs séquentiels fréquents dans un ensemble de textes. Nous avons également conçu et implémenté un système de reconnaissance d'entités nommées sur le modèle des approches symboliques classiques. Celui-ci permet d'annoter au préalable les différentes entités dites simples nécessaires à la reconnaissance des événements. Les deux méthodes pour l'extraction des événements ont montré leur efficacité lors de l'état de l'art réalisé et leurs performances ont été évaluées sur un corpus de test de notre domaine d'application (voir la section *Évaluation du travail de recherche* ci-dessous). La méthode à base de règles pourra être améliorée en tenant compte d'autres informations fournies par l'analyse syntaxique telles que la voix (passive ou active) du déclencheur, la polarité de la phrase (négative ou positive), la modalité mais aussi les phénomènes de valence multiple. L'approche à base de motifs séquentiels fréquents pourrait également tirer profit de cette analyse syntaxique en intégrant les relations de dépendance produites en tant que nouveaux items ou sous forme de contraintes. Enfin, concernant les deux approches, leur limite principale (qui est aussi celle d'autres approches de la littérature) est qu'elles réalisent l'extraction au niveau phrastique. Une granularité plus large tel que le paragraphe ou le discours pourrait permettre d'améliorer les performances de ces approches.

1.4 Un processus d'agrégation sémantique des événements

Notre contribution suivante est un processus d'agrégation sémantique des événements fondé sur une échelle de similarité qualitative et sur un ensemble de mesures spécifiques à chaque type de dimension. Nous avons tout d'abord proposé des mécanismes de normalisation des entités, adaptés à leurs natures, afin d'harmoniser formellement les différentes informations extraites. Concernant ce premier aspect, nous envisageons des améliorations telles que la désambiguïsation des dates relatives, par exemple, ou encore l'intégration au sein de notre système d'un outil de désambiguïsation sémantique des participants. Nous avons ensuite proposé un une échelle de similarité qualitative orientée utilisateur et un ensemble de calculs de similarité intégrant à la fois un modèle théorique adapté à chaque dimension et une implémentation technique employant les technologies du Web sémantique (ontologies et bases de connaissances sémantiques). Nous souhaitons poursuivre ce travail en élargissant le panel de similarités employées : notamment, des mesures de proximité ontologique plus sophistiquées ainsi que des outils de distance temporelle et spatiale et, pour la similarité agentive, une distance dédiée aux ensembles d'entités. Les similarités entre événements pourraient également provenir d'une fonction d'équivalence apprise automatiquement à partir de données annotées manuellement. Enfin, un processus d'agrégation fondé sur les graphes a été proposé afin de regrouper les mentions d'événements similaires et de permettre à l'analyste

de découvrir de nouvelles connaissances. Ce type d'agrégation possède l'avantage principal d'être intrinsèquement adapté aux traitement des bases de connaissances et ainsi aisément généralisable à d'autres silos du Web de données. Cette agrégation pourrait également être réalisée par calcul d'une similarité globale qui combinerait les différentes similarités locales. Le point délicat pour cette méthode sera alors d'estimer le poids de chaque dimension dans cette combinaison.

1.5 Évaluation du travail de recherche

Pour conclure ce bilan des contributions, nous avons proposé durant notre travail de recherche deux expérimentations destinées à évaluer deux de nos contributions. Tout d'abord, une première évaluation a porté sur notre approche mixte d'extraction automatique des événements : celle-ci a été testée et comparée à un corpus de référence (issu de la campagne d'évaluation MUC) annoté manuellement pour les besoins de l'expérimentation. Une analyse détaillée des résultats a montré de bonnes performances en comparaison avec l'état de l'art et ceci pour l'ensemble des configurations d'évaluation testées. Cette première évaluation a également pointé quelques limites de notre approche telles que l'impact de l'extraction d'entités nommées sur les performances des extracteurs d'événements. De plus, celle-ci pourrait être améliorée en exploitant davantage les informations fournies par l'analyse syntaxique en dépendance. La seconde expérimentation a constitué une première analyse qualitative dans le but d'illustrer les résultats de notre processus d'agrégation sémantique sur un jeu de données réelles. Nous avons présenté, dans un premier temps, l'implémentation d'un prototype fonctionnel couvrant la chaine complète d'agrégation des événements. Puis, la base de données *Global Terrorism Database* a été introduite ainsi qu'un sous-ensemble de celle-ci servant de corpus d'évaluation pour cette expérimentation. Les tests réalisés se sont avérés prometteurs à la fois pour ce qui est du calcul de similarité entre événements et du processus d'agrégation proposé. Le premier traitement a rapproché efficacement des mentions d'événements qui co-réfèrent et permettra ainsi de réduire la tâche de l'analyste du ROSO en lui proposant l'ensemble des liens de similarité en tant que critères supplémentaires de recherche. Puis, nous avons appliqué notre processus d'agrégation par regroupements successifs sur ce jeu de test et nous avons montré par un exemple l'utilité de cette agrégation du point de vue utilisateur. Cette seconde contribution pourra être améliorée en intégrant l'ensemble des fonctions de similarité proposées à notre prototype d'agrégation et en optimisant celui-ci pour permettre son passage à l'échelle d'un processus de veille réel.

2 Perspectives de recherche

Les travaux de recherche menés durant cette thèse ont permis de mettre en avant des perspectives d'amélioration de notre processus de capitalisation des connaissances.

La première suite à donner à nos travaux sera de mettre en place une évaluation quantitative et sur un jeu de données significatif (la base GTD par exemple) de notre processus global de capitalisation de connaissances. Pour cela, nous envisageons d'optimiser l'implémentation de notre processus d'agrégation sémantique pour passer à l'échelle de notre corpus d'évaluation.

L'extraction des événements pourra être améliorée par différentes méthodes exploitant la connaissance disponible dans le Web de données. L'ensemble des silos sémantiques liés entre eux par des équivalences et des relations ontologiques peut être exploité par divers moyens. Une autre piste d'amélioration est l'intégration d'un système d'estimation de la confiance ou qualité des extractions afin de guider soit les systèmes suivants soit l'utilisateur. De plus, il pourra être intéressant d'y ajouter d'autres méthodes performantes de l'EI comme par exemple une extraction par apprentissage statistique sous réserve de disposer d'un corpus annoté adéquat. Concernant notre approche à base de motifs séquentiels, nous pourrons diversifier le corpus d'apprentissage utilisé afin d'obtenir des patrons plus génériques et performants pour d'autres domaines ou genres de texte. Les récentes recherches visant à évaluer la qualité des extractions telles que [Habib and van Keulen, 2011], nous paraissent également à prendre en compte afin d'améliorer les performances de nos extracteurs mais aussi dans le but de réaliser une hybridation intelligente de ces différentes méthodes d'extraction.

Par ailleurs, comme dit précédemment le travail présenté ici 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 sont 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 co-référence entre entités. Nous pourrons également prendre en compte certaines dépendances entre les 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. En effet, nous n'étudions pour le moment que les événements dits ponctuels mais il sera intéressant de poursuivre ce travail en étudiant les événements duratifs et plus particulièrement les liens temporels entre événements grâce aux relations d'Allen [Allen, 1981] et à l'opérateur *convexify* défini dans le modèle TUS. Une autre perspective pourra être d'étudier tous les types de dépendance entre dimensions, notamment en explorant certaines techniques de résolution collective d'entités.

Concernant le processus d'agrégation proposé, nous envisageons d'étudier les travaux existants autour de la cotation de l'information et plus particulièrement de la fiabilité des sources et des informations. La qualité des connaissances capitalisées peut être grandement améliorée dès lors que le processus d'agrégation des informations tient compte de ces indices (voir par exemple les travaux de [Cholvy, 2007]).

Pour finir, l'interaction avec l'utilisateur constitue également une piste de recherche intéressante. [Noël, 2008] met en avant que les technologies du Web sémantique ont souvent été critiquées pour le fait que les aspects utilisateur y ont souvent été négligés. Ceux-ci proposent donc l'application des techniques de recherche exploratoire au Web sémantique et, plus particulièrement, à l'accès aux bases

de connaissances par les utilisateurs. De plus, la possibilité donnée à l'utilisateur de fusionner des fiches grâce aux suggestions d'agrégats soulèvera d'autres problématiques à explorer telles que la mise à jour et le maintien de cohérence de la base de connaissance en fonction des actions de l'analyste.

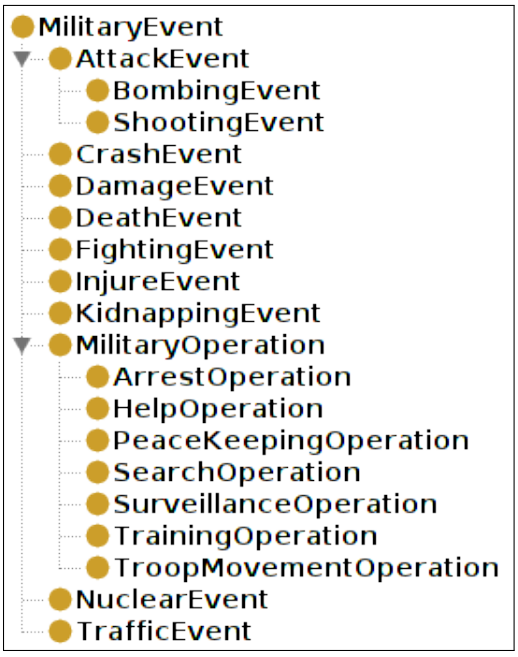
Annexes

Annexe A

WOOKIE : taxonomie des concepts

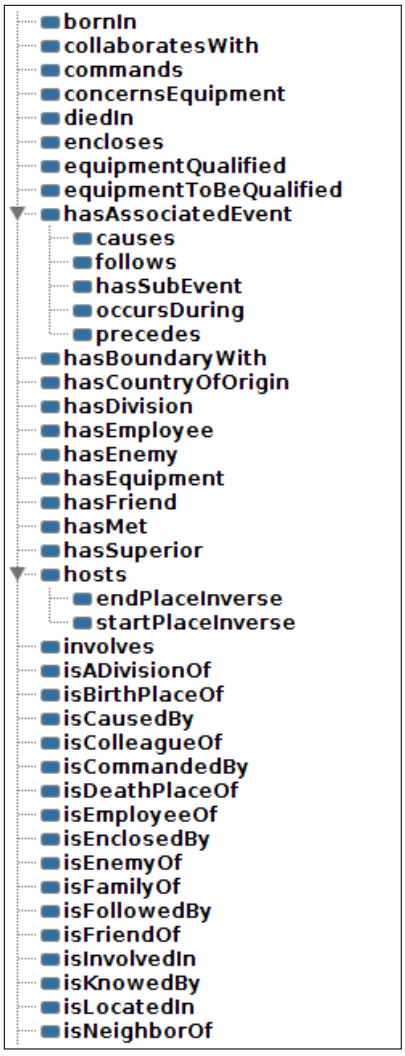
Annexe B

WOOKIE : événements spécifiques au ROSO



Annexe C

WOOKIE : relations entre concepts

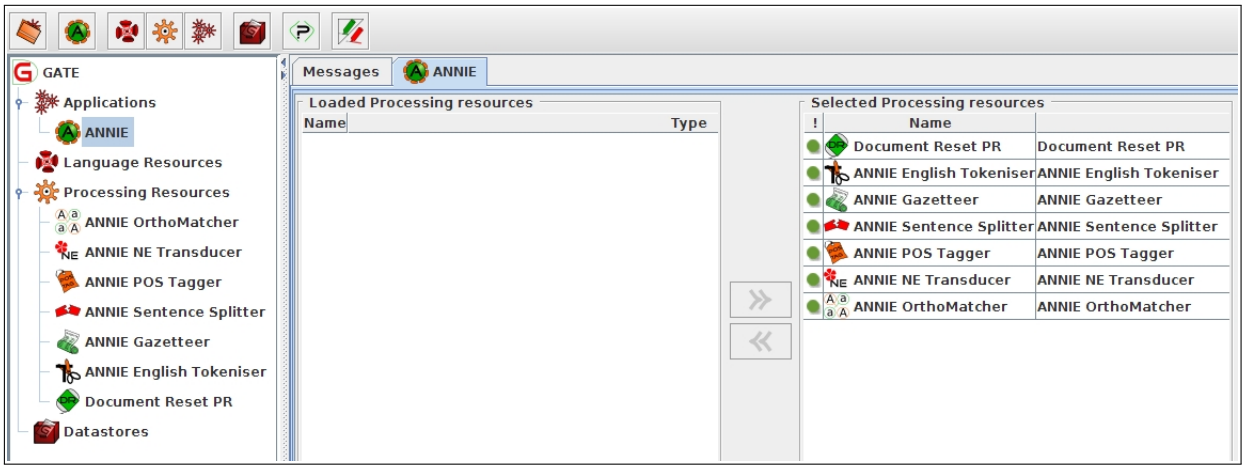


Annexe D

WOOKIE : attributs des concepts

Annexe E

GATE : exemple de chaine de traitement



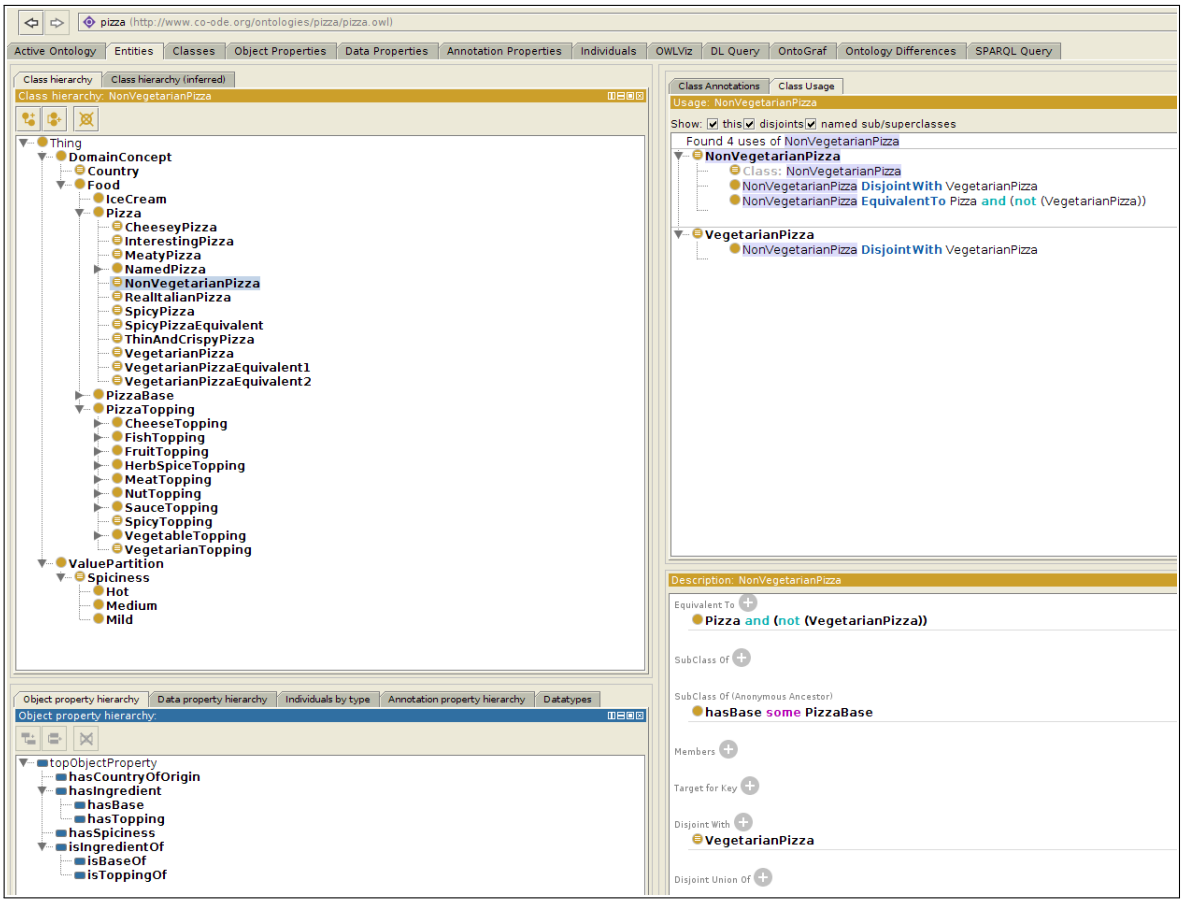
Annexe F

Gazetteer pour la détection de personnes en français

```
PDG
chef
correspondant
député
députée
directeur
directeur général
ex-président
ex-représentant
ex-responsable
gouverneur
journaliste
journalistes
leader
manager
membre
membres
pdg
porte-parole
porte-parole adjoint
président
reporter
reporters
représentant
représentant adjoint
représentant spécial
responsable
responsables
secrétaire général
témoin
trésorier
vice-président
présidente
```

Annexe G

L'ontologie-type *pizza.owl*



Annexe H

Extrait de l'ontologie *pizza.owl* au format OWL

```
<!--
    ///////////////////////////////////////////////////////////////////////
    //
    //   OWL Classes
    //
    ///////////////////////////////////////////////////////////////////////
     -->


    <!-- Class: http://www.co-ode.org/ontologies/pizza/pizza.owl#American -->

    <owl:Class rdf:about="#American">
        <rdfs:label xml:lang="pt">Americana</rdfs:label>
        <rdfs:subClassOf>
            <owl:Restriction>
                <owl:onProperty rdf:resource="#hasTopping"/>
                <owl:someValuesFrom rdf:resource="#TomatoTopping"/>
            </owl:Restriction>
        </rdfs:subClassOf>
        <rdfs:subClassOf>
            <owl:Restriction>
                <owl:onProperty rdf:resource="#hasTopping"/>
                <owl:someValuesFrom rdf:resource="#PeperoniSausageTopping"/>
            </owl:Restriction>
        </rdfs:subClassOf>
        <rdfs:subClassOf>
            <owl:Restriction>
                <owl:onProperty rdf:resource="#hasCountryOfOrigin"/>
                <owl:hasValue rdf:resource="#America"/>
            </owl:Restriction>
        </rdfs:subClassOf>
        <rdfs:subClassOf>
            <owl:Restriction>
                <owl:onProperty rdf:resource="#hasTopping"/>
                <owl:someValuesFrom rdf:resource="#MozzarellaTopping"/>
            </owl:Restriction>
```

```
    </rdfs:subClassOf>
    <rdfs:subClassOf>
        <owl:Class rdf:about="#NamedPizza"/>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="#hasTopping"/>
            <owl:allValuesFrom>
                <owl:Class>
                    <owl:unionOf rdf:parseType="Collection">
                        <owl:Class rdf:about="#MozzarellaTopping"/>
                        <owl:Class rdf:about="#PeperoniSausageTopping"/>
                        <owl:Class rdf:about="#TomatoTopping"/>
                    </owl:unionOf>
                </owl:Class>
            </owl:allValuesFrom>
        </owl:Restriction>
    </rdfs:subClassOf>
</owl:Class>

  <!--
////////////////////////////////////////////////////////////////////////////
//
//   OWL Object Properties
//
////////////////////////////////////////////////////////////////////////////
 -->


<!-- Object property: http://www.co-ode.org/ontologies/pizza/pizza.owl#hasBase -->

<owl:ObjectProperty rdf:about="#hasBase">
    <rdf:type rdf:resource="&owl;FunctionalProperty"/>
    <rdf:type rdf:resource="&owl;InverseFunctionalProperty"/>
    <owl:inverseOf>
        <owl:ObjectProperty rdf:about="#isBaseOf"/>
    </owl:inverseOf>
    <rdfs:domain>
        <owl:Class rdf:about="#Pizza"/>
    </rdfs:domain>
    <rdfs:range>
        <owl:Class rdf:about="#PizzaBase"/>
    </rdfs:range>
</owl:ObjectProperty>


   <!--
////////////////////////////////////////////////////////////////////////////
//
//   OWL Individuals
//
////////////////////////////////////////////////////////////////////////////
 -->

<!-- Individual: http://www.co-ode.org/ontologies/pizza/pizza.owl#America -->

<owl:Thing rdf:about="#America">
```

```
        <rdf:type rdf:resource="#Country"/>
    </owl:Thing>


    <!--
    ///////////////////////////////////////////////////////////////////////
    //
    //    OWL Axioms
    //
    ///////////////////////////////////////////////////////////////////////
     -->

    <owl:Class rdf:about="#LaReine">
        <owl:disjointWith>
            <owl:Class rdf:about="#Mushroom"/>
        </owl:disjointWith>
    </owl:Class>
    <owl:Class rdf:about="#Mushroom">
        <owl:disjointWith>
            <owl:Class rdf:about="#LaReine"/>
        </owl:disjointWith>
    </owl:Class>

    <owl:AllDifferent>
        <owl:distinctMembers rdf:parseType="Collection">
            <owl:Thing rdf:about="#America"/>
            <owl:Thing rdf:about="#Italy"/>
            <owl:Thing rdf:about="#Germany"/>
            <owl:Thing rdf:about="#France"/>
            <owl:Thing rdf:about="#England"/>
        </owl:distinctMembers>
    </owl:AllDifferent>
</rdf:RDF>
```

Annexe I

Exemple de document WebLab contenant des événements

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<resource xsi:type="ns3:Document" uri="weblab://SmallEnglishTest/1"
xmlns:ns3="http://weblab.ow2.org/core/1.2/model#"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <annotation uri="weblab://SmallEnglishTest/1#0-a2">
        <data xmlns:ns2="http://weblab.ow2.org/1.2/model#">
            <rdf:RDF xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
                <rdf:Description rdf:about="weblab://SmallEnglishTest/1">
                    <dc:language>en</dc:language>
                    </rdf:Description>
                </rdf:RDF>
            </data>
    </annotation>
    <mediaUnit xsi:type="ns3:Text" uri="source://xp_s78">
        <annotation uri="source://xp_s78#a0">
            <data>
                <rdf:RDF xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:dct="http://purl.org/dc/terms/"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:wlr="http://weblab.ow2.org/core/1.2/ontology/retrieval#"
xmlns:wlp="http://weblab.ow2.org/core/1.2/ontology/processing#"
xmlns:wookie="http://weblab.ow2.org/wookie#">
                    <rdf:Description rdf:about="source://xp_s78#4">
                        <wlp:refersTo
rdf:resource="http://weblab.ow2.org/wookie/instances/SearchOperation#40d72785-
2171-4372-8e61-5808b41122c3"/>
                        <wlp:refersTo
rdf:resource="http://weblab.ow2.org/wookie/instances/SearchOperation#832b376f-
c7dd-4d23-a8ea-4441027c4115"/>
                    </rdf:Description>
                    <rdf:Description rdf:about="source://xp_s78#5">
                        <wlp:refersTo
rdf:resource="http://weblab.ow2.org/wookie/instances/CrashEvent#81b9be1e-afa9-
4a84-9066-414b8021468c"/>
                        <wlp:refersTo
```

```
rdf:resource="http://weblab.ow2.org/wookie/instances/CrashEvent#5bbe4888-5445-
4896-adf0-0a20b76eed27"/>
                    </rdf:Description>
                    <rdf:Description rdf:about="source://xp_s78#a0">
                        <wlp:isProducedBy
rdf:resource="http://weblab.ow2.org/webservices/gateservice"/>
                    </rdf:Description>
                    <rdf:Description
rdf:about="http://weblab.ow2.org/wookie/instances/CrashEvent#5bbe4888-5445-4896-
adf0-0a20b76eed27">
                        <wookie:source>source://xp_s78</wookie:source>
                        <rdfs:label>incident</rdfs:label>
                        <rdf:type
rdf:resource="http://weblab.ow2.org/wookie#CrashEvent"/>
                    </rdf:Description>
                    <rdf:Description
rdf:about="http://weblab.ow2.org/wookie/instances/SearchOperation#832b376f-c7dd-
4d23-a8ea-4441027c4115">
                        <wookie:involves
rdf:resource="http://weblab.ow2.org/wookie/instances/Unit#police"/>
                        <wookie:source>source://xp_s78</wookie:source>
                        <rdfs:label>investigating</rdfs:label>
                        <rdf:type
rdf:resource="http://weblab.ow2.org/wookie#SearchOperation"/>
                    </rdf:Description>
                    <rdf:Description
rdf:about="http://weblab.ow2.org/wookie/instances/SearchOperation#40d72785-2171-
4372-8e61-5808b41122c3">
                        <wookie:source>source://xp_s78</wookie:source>
                        <rdfs:label>investigating</rdfs:label>
                        <rdf:type
rdf:resource="http://weblab.ow2.org/wookie#SearchOperation"/>
                    </rdf:Description>
                    <rdf:Description
rdf:about="http://weblab.ow2.org/wookie/instances/CrashEvent#81b9be1e-afa9-4a84-
9066-414b8021468c">
                        <wookie:involves
rdf:resource="http://weblab.ow2.org/wookie/instances/Unit#police"/>
                        <wookie:source>source://xp_s78</wookie:source>
                        <rdfs:label>incident</rdfs:label>
                        <rdf:type
rdf:resource="http://weblab.ow2.org/wookie#CrashEvent"/>
                    </rdf:Description>
                    <rdf:Description
rdf:about="http://weblab.ow2.org/wookie/instances/Person#muhammad_khan_sasoli">
                        <wlp:isCandidate>true</wlp:isCandidate>
                        <rdfs:label>Muhammad Khan Sasoli</rdfs:label>
                        <rdf:type
rdf:resource="http://weblab.ow2.org/wookie#Person"/>
                    </rdf:Description>
                    <rdf:Description
rdf:about="http://weblab.ow2.org/wookie/instances/Unit#khuzdar_press_club">
                        <wlp:isCandidate>true</wlp:isCandidate>
                        <rdfs:label>Khuzdar Press Club</rdfs:label>
                        <rdf:type
rdf:resource="http://weblab.ow2.org/wookie#Unit"/>
                    </rdf:Description>
```

```
                    <rdf:Description rdf:about="source://xp_s78#2">
                        <wlp:refersTo
rdf:resource="http://weblab.ow2.org/wookie/instances/
Person#muhammad_khan_sasoli"/>
                    </rdf:Description>
                    <rdf:Description rdf:about="source://xp_s78#3">
                        <wlp:refersTo
rdf:resource="http://weblab.ow2.org/wookie/instances/Unit#police"/>
                    </rdf:Description>
                    <rdf:Description
rdf:about="http://weblab.ow2.org/wookie/instances/Unit#police">
                        <wlp:isCandidate>true</wlp:isCandidate>
                        <rdfs:label>Police</rdfs:label>
                        <rdf:type
rdf:resource="http://weblab.ow2.org/wookie#Unit"/>
                    </rdf:Description>
                    <rdf:Description rdf:about="source://xp_s78#1">
                        <wlp:refersTo
rdf:resource="http://weblab.ow2.org/wookie/instances/Unit#khuzdar_press_club"/>
                    </rdf:Description>
                </rdf:RDF>
            </data>
        </annotation>
        <segment xsi:type="ns3:LinearSegment" start="303" end="321"
uri="source://xp_s78#1"/>
        <segment xsi:type="ns3:LinearSegment" start="386" end="406"
uri="source://xp_s78#2"/>
        <segment xsi:type="ns3:LinearSegment" start="523" end="529"
uri="source://xp_s78#3"/>
        <segment xsi:type="ns3:LinearSegment" start="533" end="546"
uri="source://xp_s78#4"/>
        <segment xsi:type="ns3:LinearSegment" start="551" end="559"
uri="source://xp_s78#5"/>
        <content>
Wednesday, December 15, 2010   E–Mail this article to a friend Printer Friendly
Version
More Sharing ServicesShare | Share on facebook Share on twitter Share on
linkedin Share on stumbleupon Share on email Share on print |

Journalist gunned down in Khuzdar

KALAT: Unidentified armed men gunned down Khuzdar Press Club president in
Khuzdar on Tuesday. According to the local police, Muhammad Khan Sasoli was on
his way home when the unidentified men gunned him down in Labour Colony. The
assailants fled from the scene. Police is investigating the incident. app
</content>
    </mediaUnit>
</resource>
```

Annexe J

Exemple de règle d'inférence au formalisme Jena

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix owl: <http://www.w3.org/2002/07/owl#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.

@prefix wookie: <http://weblab.ow2.org/wookie#>.


[Initialization:
 ->   print("//////////////␣Events␣semantic␣similarity␣rules␣//////////////////")
]



////////////// Semantic similarity ////////////////////////////////



[SemSIM_1:

        (?e1 wookie:semanticallySIM ?e2)
        <-
        (?e1 rdf:type ?c1),
        (?e2 rdf:type ?c2),
        isRelevant(?c1),
        isRelevant(?c2),
        noValue(?e1 wookie:semanticallySIM ?e2),
        notEqual(?e1,?e2),
        notEqual(?c1,?c2),
        hasSubClass(?c1,?c2)
]
```

Annexe K

Extrait d'un document du corpus d'apprentissage

{ However however RB }{ , , , }{ Canada Canada NP Place }{ continues continue VBZ }{ to to TO }{ make make VB }{ progress progress NN }{ on on IN }{ the the DT }{ six six CD }{ priorities priority NNS }{ we we PP }{ have have VBP }{ identified identify VBN }{ that that WDT }{ will will MD }{ help help VB LookupEvent }{ build build VB }{ the the DT }{ foundation foundation NN }{ for for IN }{ a a DT }{ more more RBR }{ stable stable JJ }{ Afghanistan Afghanistan NP Place }{ . . SENT }

{ The the DT }{ fifth fifth JJ }{ quarterly quarterly JJ }{ report report NN }{ highlights highlight VBZ }{ Canadian Canadian JJ }{ activity activity NN }{ in in IN }{ several several JJ }{ areas area NNS }{ : : : }{ -- : }{ Under under IN }{ a a DT }{ Canadian-supported Canadian-supported JJ }{ project project NN }{ to to TO }{ clear clear JJ }{ landmines landmine NNS }{ and and CC }{ other other JJ }{ explosives explosive NNS }{ , , , }{ training training NN LookupEvent }{ began begin VBD }{ for for IN }{ 80 @card@ CD }{ locally locally RB }{ recruited recruit VBN }{ deminers deminers NNS }{ in in IN }{ Kandahar Kandahar NP Place }{ , , , }{ and and CC }{ an an DT }{ additional additional JJ }{ 270,000 @card@ CD }{ square square JJ }{ metres metre NNS }{ of of IN }{ land land NN }{ were be VBD }{ cleared clear VBN }{ . . SENT }

{ Canada Canada NP Place }{ continues continue VBZ }{ to to TO }{ pursue pursue VB }{ its its PP$ }{ efforts effort NNS }{ to to TO }{ protect protect VB LookupEvent }{ its its PP$ }{ security security NN }{ by by IN }{ helping help VBG LookupEvent }{ the the DT }{ Afghan Afghan JJ Unit }{ government government NN Unit }{ to to TO }{ prevent prevent VB }{ Afghanistan Afghanistan NP Place }{ from from IN }{ again again RB }{ becoming become VBG }{ a a DT }{ base base NN }{ for for IN }{ terrorism terrorism NN }{ directed direct VBN }{ against against IN }{ Canada Canada NP Place }{ or or CC }{ its its PP$ }{ allies ally NNS }{ . . SENT }

Annexe L

Extrait d'un document du corpus de test

TST4–MUC4–0001

<Place>Concepcion</Place> , <Date>23 Aug 88</Date> (<Unit>Santiago Domestic Service</Unit>) –– [Report]
[<Person>Miguel Angel Valdebenito</Person>] [Text] <Unit>Police</Unit> sources have reported that
unidentified individuals <LookupEvent>planted a bomb</LookupEvent> in front of a <Place>Mormon Church</Place> in
<Place>Talcahuano District</Place> . The bomb, which <LookupEvent>exploded</LookupEvent> and <LookupEvent>caused property
damage</LookupEvent> worth 50,000 pesos, was placed at a <Place>chapel of the Church of
Jesus Christ of Latter–Day Saints</Place> located at <Place>No 3856 Gomez Carreno Street</Place> .

The shock wave destroyed a wall, the roof, and the windows of the church, but did not cause any injuries.

<Unit>Carabineros bomb squad</Unit> personnel immediately <LookupEvent>went</LookupEvent> to the location
and discovered that the bomb was made of 50 grams of an–fo [ammonium nitrate–fuel oil blasting agents] and a slow fuse.

<Unit>Carabineros special forces</Unit> soon <LookupEvent>raided</LookupEvent> a large area to try to
<LookupEvent>arrest</LookupEvent> those responsible for the attack, but they were unsuccessful.

The <Unit>police</Unit> have already informed the appropriate authorities, that is, the national prosecutor and the <Unit>Talcahuano criminal court</Unit> , of this attack.

Annexe M

Source *s12* : dépêche de presse à l'origine des événements *Event1* et *Event2*

January 24, 2010
Two U.S. Soldiers Are Among 17 Afghan Deaths
By ROD NORDLAND and SANGAR RAHIMI

KABUL, Afghanistan – At least 17 people died in four separate episodes in Afghanistan on Saturday, while a police chief was kidnapped and a provincial governor narrowly escaped assassination.

Three women and a young boy were killed when a taxi crammed with at least eight passengers tried to run an illegal Taliban checkpoint in Paktika Province, in the east, and the militants riddled the car with bullets.

Four Afghan soldiers guarding the governor of Wardak Province, just west of Kabul, were killed when the Taliban set off a hidden bomb as he traveled to a school building inspection; the governor was unharmed.

Two American soldiers were killed by an improvised explosive device in southern Afghanistan, according to a press release from the international military command here.

And seven Afghans were killed in the remote village of Qulum Balaq in Faryab Province, in northern Afghanistan, when they tried to excavate an old bomb dropped by an aircraft many years ago, according to a statement from the Interior Ministry. One person was wounded.

In addition, the police chief of Sheigal district in Kunar Province, Jamatullah Khan, and two of his officers were kidnapped while patrolling just after midnight on Saturday close to the border with Pakistan. Gen. Khalilullah Ziayee, the provincial police chief, said they were abducted "by the enemies of peace and stability in the country," the government's catch-all term for insurgents.

"We don't have any information about him yet," General Ziayee said, speaking of the police chief. He added that a search was under way.

The Taliban and common criminals often kidnap officials for ransom.

The taxicab shooting occurred as the driver was trying to take his passengers to get medical care at a nearby military base run by international forces. In addition to the three women and a boy of 5 or 6 who were killed, three other passengers were wounded, according to Mukhles Afghan, a spokesman for the provincial governor in Paktika.

The attempted assassination of the governor of Wardak, Mohammad Halim Fediyee, occurred during a trip that had been announced, leaving his convoy vulnerable.

"We were aware of the planned attack and we had already defused two bombs planted on our way," said Shahedullah Shahed, a spokesman for the governor who was traveling with him.

He said a Taliban local commander named Ahmadullah and another fighter had planted a new bomb just before the convoy crossed a culvert, detonating it under the first armored vehicle in the convoy. The blast killed four soldiers in the vehicle.

Mr. Shahed said other soldiers managed to capture the two Taliban members as they tried to flee. "This trip was an announced trip, and everybody was waiting for the governor to help them solve their problems," Mr. Shahed said. "Hundreds of tribal elders and local people were waiting to see the governor."

An Afghan employee of The New York Times in Khost Province contributed reporting.

Annexe N

Source *s3* : dépêche de presse à l'origine de l'événement *Event3*

Four kidnapped in east Afghanistan
Sun Sep 26, 2010 3:7PM
Militants have kidnapped a British woman along with three locals in eastern Afghanistan as security continues to deteriorate in the war–ravaged country.

Those abducted in the province of Kunar are reportedly employees of an American company.

Local officials have blamed the kidnapping on the Taliban but the militants have not yet claimed responsibility.

Kidnappings have recently been on the rise in Afghanistan as the security situation deteriorates to its worst levels since the 2001 US–led invasion there.

The Taliban have abducted over a dozen people across Afghanistan during the recent parliamentary elections.

This is while some 150,000 US–led foreign troops are responsible for security in the war–torn nation.

JR/AKM/MMN

Bibliographie

[Agrawal et al., 1993] Agrawal, R., Imieliński, T., and Swami, A. (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, pages 207–216, New York. ACM. 97

[Ahn, 2006] Ahn, D. (2006). The stages of event extraction. *Proceedings of the Workshop on Annotating and Reasoning about Time and Events - ARTE '06*, pages 1–8. 27, 48

[Alatrish, 2012] Alatrish, E. S. (2012). Comparison of ontology editors. *eRAF Journal on Computing*, 4 :23–38. 25

[Allen, 1981] Allen, J. F. (1981). An interval-based representation of temporal knowledge. In *Proceedings of the 7th international joint conference on Artificial intelligence - Volume 1*, pages 221–226, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. 136

[Allen, 1983] Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11) :832–843. 33

[Allen, 1991] Allen, J. F. (1991). Time and time again : The many ways to represent time. *Journal of Intelligent Systems*, 6(4) :341–355. 33

[Allen and Ferguson, 1994] Allen, J. F. and Ferguson, G. (1994). Actions and events in interval temporal logic. *Journal of Logic and Computation*, 4 :531–579. 29

[Aone et al., 1998] Aone, C., Halverson, L., Hampton, T., and Ramos-Santacruz, M. (1998). SRA : Description of the IE2 system used for MUC-7. In *Proceedings Seventh Message Understanding Conference (MUC-7)*, Fairfax, VA. 48

[Aone and Ramos-Santacruz, 2000] Aone, C. and Ramos-Santacruz, M. (2000). REES : A large-scale relation and event extraction system. In *ANLP*, pages 76–83. 48

[Appelt, 1999] Appelt, D. E. (1999). Introduction to information extraction. *AI Commun.*, 12(3) :161–172. 56

[Appelt et al., 1995] Appelt, D. E., Hobbs, J. R., Bear, J., Israel, D., Kameyama, M., Martin, D., Myers, K., and Tyson, M. (1995). SRI International FASTUS system : MUC-6 test results and analysis. In *Proceedings of the 6th conference on Message understanding*, MUC6 '95, pages 237–248, Stroudsburg, PA, USA. Association for Computational Linguistics. 48

[Appelt and Onyshkevych, 1998] Appelt, D. E. and Onyshkevych, B. (1998). The common pattern specification language. In *Proceedings of a workshop on held at Baltimore, Maryland : October 13-15, 1998*, TIPSTER '98, pages 23–30, Stroudsburg, PA, USA. Association for Computational Linguistics. 44

[Augenstein et al., 2012] Augenstein, I., Padó, S., and Rudolph, S. (2012). Lodifier : generating linked data from unstructured text. In *Proceedings of the 9th international conference on The Semantic Web : research and applications*, ESWC'12, pages 210–224, Berlin, Heidelberg. Springer-Verlag. 65

[Bachmair and Ganzinger, 2001] Bachmair, L. and Ganzinger, H. (2001). Resolution theorem proving. In *Handbook of Automated Reasoning*, pages 19–99. Elsevier and MIT Press. 22

[Bagga and Baldwin, 1999] Bagga, A. and Baldwin, B. (1999). Cross-document event coreference : Annotations, experiments, and observations. In *In Proc. ACL-99 Workshop on Coreference and Its Applications*, pages 1–8. 64, 66

[Balmisse, 2002] Balmisse, G. (2002). *Gestion des connaissances. Outils et applications du knowledge management*. Vuibert. 18

[Baumgartner and Retschitzegger, 2006] Baumgartner, N. and Retschitzegger, W. (2006). A survey of upper ontologies for situation awareness. *Proc. of the 4th IASTED International Conference on Knowledge Sharing and Collaborative Engineering, St. Thomas, US VI*, pages 1–9+. 36

[Béchet et al., 2012] Béchet, N., Cellier, P., Charnois, T., and Crémilleux, B. (2012). Discovering linguistic patterns using sequence mining. In *CICLing (1)*, pages 154–165. 97, 98

[Benjelloun et al., 2006] Benjelloun, O., Garcia-Molina, H., Kawai, H., Larson, T. E., Menestrina, D., Su, Q., Thavisomboon, S., and Widom, J. (2006). Generic entity resolution in the serf project. *IEEE Data Eng. Bull.*, 29(2) :13–20. 63

[Berners-Lee et al., 2001] Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5) :34–43. 19

[Besançon et al., 2010] Besançon, R., de Chalendar, G., Ferret, O., Gara, F., Mesnard, O., Laïb, M., and Semmar, N. (2010). LIMA : A multilingual framework for linguistic analysis and linguistic resources development and evaluation. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA). 49

[Besançon et al., 2011] Besançon, R., Ferret, O., and Jean-Louis, L. (2011). Construire et évaluer une application de veille pour l'information sur les événements sismiques. In *CORIA*, pages 287–294. 53

[Best and Cumming, 2007] Best, R. and Cumming, A. (2007). Open source intelligence (osint) : Issues for congress. Rl 34270, Congressional Research Service. 3

[Bhattacharya and Getoor, 2007] Bhattacharya, I. and Getoor, L. (2007). Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data*, 1(1) :5–es. 63

[Bilenko et al., 2003] Bilenko, M., Mooney, R. J., Cohen, W. W., Ravikumar, P., and Fienberg, S. E. (2003). Adaptive name matching in information integration. *IEEE Intelligent Systems*, 18(5) :16–23. 65, 66

[Bloch, 2005] Bloch, I. (2005). Fusion d'informations numériques : panorama méthodologique. In *Journées Nationales de la Recherche en Robotique 2005*, pages 79–88, Guidel, France. 62

[Bond et al., 2003] Bond, D., Bond, J., Oh, C., Jenkins, J. C., and Taylor, C. L. (2003). Integrated Data for Events Analysis (IDEA) : An Event Typology for Automated Events Data Development. *Journal of Peace Research*, 40(6) :733–745. 36

[Bontcheva et al., 2002] Bontcheva, K., Dimitrov, M., Maynard, D., Tablan, V., and Cunningham, H. (2002). Shallow Methods for Named Entity Coreference Resolution. In *TALN 2002*. 46

[Borsje et al., 2010] Borsje, J., Hogenboom, F., and Frasincar, F. (2010). Semi-automatic financial events discovery based on lexico-semantic patterns. *Int. J. Web Eng. Technol.*, 6(2) :115–140. 53

[Boury-Brisset, 2003] Boury-Brisset, A.-C. (2003). Ontological approach to military knowledge modeling and management. In *NATO RTO Information Systems Technology Symposium (RTO MP IST 040)*, Prague. 36

[Bowman et al., 2001] Bowman, M., Lopez, A. M., and Tecuci, G. (2001). Ontology development for military applications. In *Proceedings of the Thirty-ninth Annual ACM Southeast Conference*. ACM Press. 36

[Brill, 1992] Brill, E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the third conference on Applied natural language processing*, ANLC '92, pages 152–155, Stroudsburg, PA, USA. Association for Computational Linguistics. 86

[Bundschus et al., 2008] Bundschus, M., Dejori, M., Stetter, M., Tresp, V., and Kriegel, H.-P. (2008). Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics*, 9(1) :1–14. 53

[Buscaldi, 2010] Buscaldi, D. (2010). *Toponym Disambiguation in Information Retrieval*. PhD thesis, Universidad Politecnica de Valencia. 103

[Califf and Mooney, 2003] Califf, M. E. and Mooney, R. J. (2003). Bottom-Up Relational Learning of Pattern Matching Rules for Information Extraction. *J. Mach. Learn. Res.*, 4 :177–210. 43, 54

[Capet et al., 2011] Capet, P., Delavallade, T., Généreux, M., Poibeau, T., Sándor, Á., and Voyatzi, S. (2011). Un système de détection de crise basé sur l'extraction automatique d'événements. In et P. Hoogstoel, M. C., editor, *Sémantique et multimodalité en analyse de l'information*, pages 293–313. Lavoisier. 42

[Capet et al., 2008] Capet, P., Delavallade, T., Nakamura, T., Sandor, A., Tarsitano, C., and Voyatzi, S. (2008). A risk assessment system with automatic extraction of event types. In Shi, Z., Mercier-Laurent, E., and Leake, D., editors, *Intelligent Information Processing IV*, volume 288 of *IFIP – The International Federation for Information Processing*, pages 220–229. Springer US. 53

[Caron et al., 2012] Caron, C., Guillaumont, J., Saval, A., and Serrano, L. (2012). Weblab : une plateforme collaborative dédiée à la capitalisation de connaissances. In *Extraction et gestion des connaissances (EGC'2012)*, Bordeaux, France. 77

[Casati and Varzi, 1997] Casati, R. and Varzi, A. (1997). Fifty years of events : an annotated bibliography 1947 to 1997. `http://www.pdcnet.org/pages/Products/electronic/eventsbib.htm`. 26

[Cellier and Charnois, 2010] Cellier, P. and Charnois, T. (2010). Fouille de données séquentielle d'itemsets pour l'apprentissage de patrons linguistiques. In *Traitement Automatique des Langues Naturelles (short paper)*. 58

[Cellier et al., 2010] Cellier, P., Charnois, T., and Plantevit, M. (2010). Sequential patterns to discover and characterise biological relations. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 6008 of *Lecture Notes in Computer Science*, pages 537–548. Springer Berlin Heidelberg. 47

[Ceri et al., 1989] Ceri, S., Gottlob, G., and Tanca, L. (1989). What you always wanted to know about datalog (and never dared to ask). *IEEE Transactions on Knowledge and Data Engineering*, 1(1) :146–166. 44

[Charlet et al., 2004] Charlet, J., Bachimont, B., and Troncy, R. (2004). Ontologies pour le Web sémantique. In *Revue I3, numéro Hors Série «Web sémantique»*. Cépaduès. 21, 25

[Charlot and Lancini, 2002] Charlot, J.-M. and Lancini, A. (2002). De la connaissance aux systèmes d'information supports. In Rowe, F., editor, *Faire de la recherche en systèmes d'information*, pages 139–145. Vuibert FNEGE. 18

[Charnois et al., 2009] Charnois, T., Plantevit, M., Rigotti, C., and Cremilleux, B. (2009). Fouille de données séquentielles pour l'extraction d'information dans les textes. *Revue Traitement Automatique des Langues (TAL)*, 50(3) :59–87. 45

[Charton et al., 2011] Charton, E., Gagnon, M., and Ozell, B. (2011). Génération automatique de motifs de détection d'entités nommées en utilisant des contenus encyclopédiques. In *18e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2011)*, Montpellier. Association pour le Traitement Automatique des Langues (ATALA). 45

[Chasin, 2010] Chasin, R. (2010). Event and temporal information extraction towards timelines of wikipedia articles. In *UCCS REU 2010*, pages 1–9. Massachusetts Institute of Technology. 54

[Chau and Xu, 2012] Chau, M. and Xu, J. (2012). Business intelligence in blogs : understanding consumer interactions and communities. *MIS Q.*, 36(4) :1189–1216. 53

[Chau et al., 2002] Chau, M., Xu, J. J., and Chen, H. (2002). Extracting meaningful entities from police narrative reports. In *Proceedings of the 2002 annual national conference on Digital government research*, dg.o '02, pages 1–5. Digital Government Society of North America. 54

[Chaudet, 2004] Chaudet, H. (2004). Steel : A spatio-temporal extended event language for tracking epidemic spread from outbreak reports. In *In U. Hahn (Ed.), Proceedings of KR-MED 2004, First International Workshop on Formal Biomedical Knowledge Representation*. 53

[Chen and Ji, 2009] Chen, Z. and Ji, H. (2009). Graph-based event coreference resolution. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, TextGraphs-4, pages 54–57, Stroudsburg, PA, USA. Association for Computational Linguistics. 66, 67

[Chieu, 2003] Chieu, H. L. (2003). Closing the gap : Learning-based information extraction rivaling knowledge-engineering methods. In *In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 216–223. 48

[Chisholm, 1970] Chisholm, R. (1970). Events and propositions. *Noûs*, 4(1) :15–24. 26

[Chiticariu et al., 2010] Chiticariu, L., Krishnamurthy, R., Li, Y., Reiss, F., and Vaithyanathan, S. (2010). Domain adaptation of rule-based annotators for named-entity recognition tasks. In *In EMNLP (To appear*. 58

[Cholvy, 2007] Cholvy, L. (2007). Modelling information evaluation in fusion. In *FUSION*, pages 1–6. 136

[Ciravegna, 2001] Ciravegna, F. (2001). Adaptive information extraction from text by rule induction and generalisation. In *Proceedings of the 17th international joint conference on Artificial intelligence - Volume 2*, IJCAI'01, pages 1251–1256, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. 45, 54

[Crié, 2003] Crié, D. (2003). De l'extraction des connaissances au Knowledge Management. *Revue française de gestion*, 29(146) :59–79. 18

[Cucerzan, 2007] Cucerzan, S. (2007). Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic. Association for Computational Linguistics. 105

[Culotta et al., 2006] Culotta, A., Kristjansson, T., McCallum, A., and Viola, P. (2006). Corrective feedback and persistent learning for information extraction. *Artificial Intelligence*, 170(14–15) :1101 – 1122. 58

[Cunningham et al., 2002] Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). GATE : A framework and graphical development environment for robust nlp tools and applications. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA*. 51, 84

[Cunningham et al., 2000] Cunningham, H., Maynard, D., and Tablan, V. (2000). JAPE : a Java Annotation Patterns Engine (Second Edition). Technical Report Technical Report CS–00–10, of Sheffield, Department of Computer Science. 44

[Daille et al., 2000] Daille, B., Fourour, N., and Morin, E. (2000). Catégorisation des noms propres : une étude en corpus. In *Cahiers de Grammaire - Sémantique et Corpus*, volume 25, pages 115–129. Université de Toulouse-le-Mirail. 43

[Daumé et al., 2010] Daumé, III, H., Kumar, A., and Saha, A. (2010). Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, DANLP 2010, pages 53–59, Stroudsburg, PA, USA. Association for Computational Linguistics. 58

[Davidson, 1967] Davidson, D. (1967). The logical form of action sentences. In Rescher, N., editor, *The Logic of Decision and Action*. University of Pittsburgh Press, Pittsburgh. 26

[Davidson, 1969] Davidson, D. (1969). The individuation of events. In Rescher, N., editor, *Essays in honor of Carl G. Hempel*, pages 216–234. D. Reidel, Dordrecht. reprinted in Davidson, Essays on Actions and Events. 27

[De Marneffe and Manning, 2008] De Marneffe, M.-C. and Manning, C. D. (2008). Stanford typed dependencies manual. Technical report, Stanford University. 95

[Desclés, 1990] Desclés, J.-P. (1990). "State, event, process, and topology". *General Linguistics*, 29(3) :159–200. 26

[Desodt-Lebrun, 1996] Desodt-Lebrun, A.-M. (1996). Fusion de données. In *Techniques de l'ingénieur Automatique avancée*, number 12 in 96, pages 1–9. Editions Techniques de l'Ingénieur. 62

[Dey et al., 1998] Dey, D., Sarkar, S., and De, P. (1998). A probabilistic decision model for entity matching in heterogeneous databases. *Management Science*, 44(10) :1379–1395. 63

[Dong and Pei, 2007] Dong, G. and Pei, J. (2007). *Sequence Data Mining*, volume 33 of *Advances in Database Systems*. Kluwer. 99

[Dong et al., 2005] Dong, X., Halevy, A., and Madhavan, J. (2005). Reference reconciliation in complex information spaces. *Proceedings of the 2005 ACM SIGMOD international conference on Management of data - SIGMOD '05*, page 85. 63

[Dredze et al., 2010] Dredze, M., McNamee, P., Rao, D., Gerber, A., and Finin, T. (2010). Entity disambiguation for knowledge base population. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 277–285, Stroudsburg, PA, USA. Association for Computational Linguistics. 64, 65

[Drozdzynski et al., 2004] Drozdzynski, W., Krieger, H.-U., Piskorski, J., Schäfer, U., and Xu, F. (2004). Shallow processing with unification and typed feature structures — foundations and applications. *Künstliche Intelligenz*, 1 :17–23. 49

[Elmagarmid et al., 2007] Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. (2007). Duplicate record detection : A survey. *IEEE Trans. on Knowl. and Data Eng.*, 19(1) :1–16. 63, 65, 66

[Enjalbert, 2008] Enjalbert, P. (2008). « Préface ». In *Plate-formes pour le traitement automatique des langues*, volume 49 of *Revue interntaionale Traitement Automatique des Langues*, chapter 2, pages 7–10. ATALA. 50

[Etzioni et al., 2005] Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2005). Unsupervised named-entity extraction from the web : An experimental study. *Artif. Intell.*, 165(1) :91–134. 43

[Fellegi and Sunter, 1969] Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64 :1183–1210. 63

[Fensel et al., 2001] Fensel, D., van Harmelen, F., Horrocks, I., McGuinness, D. L., and Patel-Schneider, P. F. (2001). OIL : An ontology infrastructure for the semantic web. *IEEE Intelligent Systems*, 16(2) :38–45. 22

[Ferré, 2007] Ferré, S. (2007). CAMELIS : Organizing and Browsing a Personal Photo Collection with a Logical Information System. In Diatta, J., Eklund, P., and Liquière, M., editors, *Int. Conf. Concept Lattices and Their Applications*, volume 331, pages 112–123, Montpellier, France. 98, 99

[Fialho et al., 2010] Fialho, A., Troncy, R., Hardman, L., Saathoff, C., and Scherp, A. (2010). What's on this evening ? Designing user support for event-based annotation and exploration of media. In *EVENTS 2010, 1st International Workshop on EVENTS - Recognising and tracking events on the Web and in real life, May 4, 2010, Athens, Greece*, Athens, GRÈCE. 29

[Finin et al., 2009] Finin, T., Syed, Z., Mayfield, J., McNamee, P., and Piatko, C. (2009). Using Wikitology for Cross-Document Entity Coreference Resolution. In *Proceedings of the AAAI Spring Symposium on Learning by Reading and Learning to Read*. AAAI Press. 46, 109

[Fisher et al., 2005] Fisher, M., Gabbay, D., and Vila, L. (2005). *Handbook of Temporal Reasoning in Artificial Intelligence*. Foundations of Artificial Intelligence. Elsevier Science. 76

[Fleischman and Hovy, 2002] Fleischman, M. and Hovy, E. (2002). Fine grained classification of named entities. *Proceedings of the 19th international conference on Computational linguistics* -, 1 :1–7. 58

[Fourour, 2002] Fourour, N. (2002). Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français. *Actes de la 9ème Conférence Nationale sur le Traitement Automatique des Langues Naturelles (TALN 2001)*, 1 :265–274. 45

[François et al., 2007] François, J., Le Pesant, D., and Leeman, D. (2007). Présentation de la classification des verbes français de jean dubois et françoise dubois-charlier. *Langue Française*, 153(153) :3–32. 96

[Friburger, 2006] Friburger, N. (2006). « Linguistique et reconnaissance automatique des noms propres ». *Meta : journal des traducteurs / Meta : Translators' Journal*, 51(4) :637–650. 44

[Fundel et al., 2007] Fundel, K., Küffner, R., Zimmer, R., and Miyano, S. (2007). Relex–relation extraction using dependency parse trees. *Bioinformatics*, 23. 46

[Garbin and Mani, 2005] Garbin, E. and Mani, I. (2005). Disambiguating toponyms in news. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. 103

[Genesereth, 1991] Genesereth, M. R. (1991). Knowledge interchange format. In *KR*, pages 599–600. 22

[Giroux et al., 2008] Giroux, P., Brunessaux, S., Brunessaux, S., Doucy, J., Dupont, G., Grilheres, B., Mombrun, Y., and Saval, A. (2008). Weblab : An integration infrastructure to ease the development of multimedia processing applications. *ICSSEA*. 6

[Goujon, 2002] Goujon, B. (2002). Annotation d'événements dans les textes pour la veille stratégique. *Event (London)*. 53

[Grishman et al., 2002a] Grishman, R., Huttunen, S., and Yangarber, R. (2002a). Information extraction for enhanced access to disease outbreak reports. *Journal of biomedical informatics*, 35(4) :236–46. 53

[Grishman et al., 2002b] Grishman, R., Huttunen, S., and Yangarber, R. (2002b). Real-time event extraction for infectious disease outbreaks. *Proceedings of the second international conference on Human Language Technology Research* -, pages 366–369. 48

[Grishman and Sundheim, 1996] Grishman, R. and Sundheim, B. (1996). Message understanding conference-6 : a brief history. In *Proceedings of the 16th conference on Computational linguistics - Volume 1*, pages 466–471, Morristown, NJ, USA. Association for Computational Linguistics. 41

[Gruber, 1993] Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5(2) :199–220. 21

[Guarino, 1998] Guarino, N. (1998). Formal ontology and information systems. In *Proceedings of Formal Ontology in Information System*, pages 3–15. IOS Press. 22

[Haase et al., 2008] Haase, P., Lewen, H., Studer, R., Tran, D. T., Erdmann, M., d'Aquin, M., and Motta, E. (2008). The NeOn Ontology Engineering Toolkit. In *WWW 2008 Developers Track*. 25

[Habib and van Keulen, 2011] Habib, M. B. and van Keulen, M. (2011). Improving named entity disambiguation by iteratively enhancing certainty of extraction. Technical Report TR-CTIT-11-29, Centre for Telematics and Information Technology University of Twente, Enschede. 136

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software : an update. *SIGKDD Explor. Newsl.*, 11(1) :10–18. 51

[Hasegawa et al., 2004] Hasegawa, T., Sekine, S., and Grishman, R. (2004). Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics. 47

[Hayes, 1995] Hayes, P. (1995). A catalog of temporal theories. Technical report, University of Illinois. Tech report UIUC-BI-AI-96-01. 33

[Hecking, 2003] Hecking, M. (2003). Information extraction from battlefield reports. *Proceedings of the 8thInternational Command and Control Research and Technology Symposium (ICCRTS)*. 53

[Higginbotham et al., 2000] Higginbotham, J., Pianesi, F., and Varzi, A. (2000). *Speaking of Events*. Oxford University Press. 26

[Hobbs et al., 1997] Hobbs, J. R., Appelt, D. E., Bear, J., Israel, D. J., Kameyama, M., Stickel, M. E., and Tyson, M. (1997). FASTUS : A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. *CoRR*, cmp-lg/9705013. 44

[Hobbs and Riloff, 2010] Hobbs, J. R. and Riloff, E. (2010). Information extraction. In *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL. 40, 43

[Hogenboom et al., 2011] Hogenboom, F., Frasincar, F., Kaymak, U., and de Jong, F. (2011). An Overview of Event Extraction from Text. In van Erp, M., van Hage, W. R., Hollink, L., Jameson, A., and Troncy, R., editors, *Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011) at Tenth International Semantic Web Conference (ISWC 2011)*, volume 779 of *CEUR Workshop Proceedings*, pages 48–57. CEUR-WS.org. 42, 56

[Horrocks, 2002] Horrocks, I. (2002). daml+oil : a description logic for the semantic web. *IEEE Data Engineering Bulletin*, 25 :4–9. 22

[Huffman, 1995] Huffman, S. (1995). Learning information extraction patterns from examples. In *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, pages 246–260. Springer. 49

[Humphreys et al., 1997] Humphreys, K., Gaizauskas, R., and Azzam, S. (1997). Event coreference for information extraction. In *Proceedings of the ACL-97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, Madrid, Spain. 66

[Inyaem et al., 2010a] Inyaem, U., Haruechaiyasak, C., Meesad, P., and Tran, D. (2010a). Terrorism event classification using fuzzy inference systems. *CoRR*, abs/1004.1772. 53

[Inyaem et al., 2010b] Inyaem, U., Meesad, P., Haruechaiyasak, C., and Tran, D. (2010b). Construction of fuzzy ontology-based terrorism event extraction. In *WKDD*, pages 391–394. IEEE Computer Society. 36

[Ireson et al., 2005] Ireson, N., Ciravegna, F., Califf, M. E., Freitag, D., Kushmerick, N., and Lavelli, A. (2005). Evaluating machine learning for information extraction. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, pages 345–352, New York, NY, USA. ACM. 43

[Isozaki and Kazawa, 2002] Isozaki, H. and Kazawa, H. (2002). Efficient support vector classifiers for named entity recognition. In *In Proceedings of the 19th International Conference on Computational Linguistics (COLING'02*, pages 390–396. 45

[Ittoo et al., 2006] Ittoo, A., Zhang, Y., and Jiao, J. (2006). A text mining-based recommendation system for customer decision making in online product customization. In *Management of Innovation and Technology, 2006 IEEE International Conference on*, volume 1, pages 473–477. 54

[Jansche and Abney, 2002] Jansche, M. and Abney, S. P. (2002). Information extraction from voicemail transcripts. In *In Proc. Conference on Empirical Methods in NLP*. 53

[Jarrar and Meersman, 2009] Jarrar, M. and Meersman, R. (2009). Ontology engineering — the DOGMA approach. In Dillon, T. S., Chang, E., Meersman, R., and Sycara, K., editors, *Advances in Web Semantics I*, pages 7–34. Springer-Verlag, Berlin, Heidelberg. 22

[Jason et al., 2004] Jason, R. S., Crawford, J., Kephart, J., and Leiba, B. (2004). Spamguru : An enterprise anti-spam filtering system. In *In Proceedings of the First Conference on E-mail and Anti-Spam*, page 2004. 54

[Jean-Louis et al., 2012] Jean-Louis, L., Romaric, B., and Ferret, O. (2012). Une méthode d'extraction d'information fondée sur les graphes pour le remplissage de formulaires. In *Actes de la 19e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2012)*, pages 29–42, Grenoble, France. 49

[Ji, 2010] Ji, H. (2010). Challenges from information extraction to information fusion. In *Proceedings of the 23rd International Conference on Computational Linguistics : Posters*, COLING '10, pages 507–515, Stroudsburg, PA, USA. Association for Computational Linguistics. 62

[Ji et al., 2009] Ji, H., Grishman, R., Chen, Z., and Gupta, P. (2009). Cross-document Event Extraction and Tracking : Task, Evaluation, Techniques and Challenges. *Society*, pages 166–172. 64

[Jonasson, 1994] Jonasson, K. (1994). *Le Nom Propre, Constructions et interprétations*. Champs linguistiques. Duculot. 44

[Khrouf and Troncy, 2012] Khrouf, H. and Troncy, R. (2012). Réconcilier les événements dans le web de données. In *Actes de IC2011*, pages 723–738, Chambéry, France. 66, 67

[Kifer et al., 1995] Kifer, M., Lausen, G., and Wu, J. (1995). Logical foundations of object-oriented and frame-based languages. *J. ACM*, 42(4) :741–843. 22

[Kim, 1973] Kim, J. (1973). Causation, nomic subsumption, and the concept of event. *Journal of Philosophy*, 70(8) :217–236. 26

[Krieg-Planque, 2009] Krieg-Planque, A. (2009). *A propos des noms propres d'événement*, volume 11, pages 77–90. Les carnets du Cediscor. 27

[Kripke, 1980] Kripke, S. (1980). *Naming and Necessity*. Harvard University Press. 41

[Ladkin, 1987] Ladkin, P. (1987). The logic of time representation. phdphd, University of California, Berkeley. 33, 76

[Lafferty et al., 2001] Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. 45

[LaFree and Dugan, 2007] LaFree, G. and Dugan, L. (2007). Introducing the Global Terrorism Database. *Terrorism and Political Violence*, 19(2) :181–204. 123

[Largeron et al., 2009] Largeron, C., Kaddour, B., and Fernandez, M. (2009). Softjaccard : une mesure de similarité entre ensembles de chaînes de caratères pour l'unification d'entités nommées. In Ganascia, J.-G. and Gançarski, P., editors, *EGC*, volume RNTI-E-15 of *Revue des Nouvelles Technologies de l'Information*, pages 443–444. Cépaduès-Éditions. 112

[Lavelli et al., 2004] Lavelli, A., Califf, M. E., Ciravegna, F., Freitag, D., Giuliano, C., Kushmerick, N., and Romano, L. (2004). IE evaluation : Criticisms and recommendations. In *In AAAI-2004 Workshop on Adaptive Text Extraction and Mining*. 58

[Lawrence et al., 1999] Lawrence, S., Giles, C. L., and Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *IEEE COMPUTER*, 32(6) :67–71. 54

[LDC, 2005] LDC, L. D. C. (2005). *ACE (Automatic Content Extraction) : English annotation guidelines for events*, version 5.4.3 2005.07.01 edition edition. 28

[Lee et al., 2012] Lee, H., Recasens, M., Chang, A., Surdeanu, M., and Jurafsky, D. (2012). Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 489–500, Stroudsburg, PA, USA. Association for Computational Linguistics. 66

[Lejeune et al., 2010] Lejeune, G., Doucet, A., Yangarber, R., and Lucas, N. (2010). Filtering news for epidemic surveillance : towards processing more languages with fewer resources. In *CLIA/COLING*, pages 3–10. 53

[Ligeza and Bouzid, 2008] Ligeza, A. and Bouzid, M. (2008). Temporal specifications with xtus. a hierarchical algebraic approach. In Cotta, C., Reich, S., Schaefer, R., and Ligęza, A., editors, *Knowledge-Driven Computing*, volume 102 of *Studies in Computational Intelligence*, pages 133–148. Springer Berlin / Heidelberg. 76

[Liu et al., 2008] Liu, M., Liu, Y., Xiang, L., Chen, X., and Yang, Q. (2008). Extracting key entities and significant events from online daily news. In *IDEAL*, pages 201–209. 53

[Llorens Martínez, 2011] Llorens Martínez, H. (2011). *A semantic approach to temporal information processing*. PhD thesis, Universidad de Alicante. 103

[Lofi et al., 2012] Lofi, C., Selke, J., and Balke, W.-T. (2012). Information extraction meets crowdsourcing : A promising couple. *Datenbank-Spektrum*, 12(2) :109–120. 58

[Luberg et al., 2012] Luberg, A., Järv, P., and Tammet, T. (2012). Information extraction for a tourist recommender system. In Fuchs, M., Ricci, F., and Cantoni, L., editors, *Information and Communication Technologies in Tourism 2012*, pages 332–343. Springer Vienna. 54

[Mann and Yarowsky, 2003] Mann, G. S. and Yarowsky, D. (2003). Unsupervised personal name disambiguation. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 33–40, Stroudsburg, PA, USA. Association for Computational Linguistics. 105

[Mannes and Golbeck, 2005] Mannes, A. and Golbeck, J. (2005). Building a terrorism ontology. In *ISWC Workshop on Ontology Patterns for the Semantic Web*. 36

[Maurel et al., 2011] Maurel, D., Friburger, N., Antoine, J.-Y., Eshkol, I., and Nouvel, D. (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées. *Traitement Automatique des Langues*, 52(1) :69–96. 44

[McCallum, 2005] McCallum, A. (2005). Information extraction : Distilling structured data from unstructured text. *Queue*, 3(9) :48–57. 51

[McDermott, 1982] McDermott, D. (1982). A temporal logic for reasoning about processes and plans*. *Cognitive Science*, 6(2) :101–155. 33

[McDonald, 1996] McDonald, D. D. (1996). Internal and external evidence in the identification and semantic categorization of proper names. In Boguraev, B. and Pustejovsky, J., editors, *Corpus processing for lexical acquisition*, pages 21–39. MIT Press, Cambridge, MA, USA. 44

[Mihalcea and Csomai, 2007] Mihalcea, R. and Csomai, A. (2007). Wikify ! : linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 233–242, New York, NY, USA. ACM. 58

[Mikheev, 1999] Mikheev, A. (1999). A knowledge-free method for capitalized word disambiguation. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* -, pages 159–166. 89

[Mikheev et al., 1999] Mikheev, A., Moens, M., and Grover, C. (1999). Named entity recognition without gazetteers. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, EACL'99, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics. 45

[Milne and Witten, 2008] Milne, D. and Witten, I. H. (2008). Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 509–518, New York, NY, USA. ACM. 58, 65

[Minard, 2008] Minard, A.-L. (2008). Etat de l'art des ontologies d'objets géographiques. Master's thesis, Laboratoire COGIT (IGN). 35

[Mintz et al., 2009] Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual*

Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 2 - Volume 2, ACL '09, pages 1003–1011, Stroudsburg, PA, USA. Association for Computational Linguistics. 47

[Mizoguchi, 2003a] Mizoguchi, R. (2003a). Tutorial on ontological engineering : Part 1 : Introduction to ontological engineering. *New Generation Comput.*, 21(4) :365–384. 22, 78

[Mizoguchi, 2003b] Mizoguchi, R. (2003b). Tutorial on ontological engineering : Part 2 : Ontology development, tools and languages. *New Generation Comput.*, 22(1) :61–96. 22

[Montague, 1969] Montague, R. (1969). On the nature of certain philosophical entities. *The Monist*, 53(2) :159–194. 26

[Moreau et al., 2008] Moreau, E., Yvon, F., and Cappé, O. (2008). Robust similarity measures for named entities matching. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 593–600, Stroudsburg, PA, USA. Association for Computational Linguistics. 66

[Muller and Tannier, 2004] Muller, P. and Tannier, X. (2004). Annotating and measuring temporal relations in texts. In *Coling 2004 , Genève,*, pages 50–56. Association for Computational Linguistics. 46

[Muslea, 1999] Muslea, I. (1999). Extraction Patterns for Information Extraction Tasks : A Survey. *Proc. AAAI-99 Workshop Machine Learning for Information Extraction*, pages 1–6. 43

[Nadeau and Sekine, 2007] Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1) :3–26. Publisher : John Benjamins Publishing Company. 41, 43, 44, 58

[Nakamura-Delloye and Villemonte De La Clergerie, 2010] Nakamura-Delloye, Y. and Villemonte De La Clergerie, É. (2010). Exploitation de résultats d'analyse syntaxique pour extraction semi-supervisée des chemins de relations. In *TALN 2010*, page taln2010_submission_164, Montréal, Canada. 46

[NATO, 2001] NATO (2001). The NATO Military Intelligence Data Exchange Standard AINTP-3(A). Technical report, NATO. 36

[NATO, 2007] NATO (2007). Joint c3 information exchange data model - jc3iedm. Technical report, NATO. 36

[Naughton et al., 2006] Naughton, M., Kushmerick, N., and Carthy, J. (2006). Event extraction from heterogeneous news sources. In *Proc. Workshop Event Extraction and Synthesis*. American Nat. Conf. Artificial Intelligence. 41, 66

[Neches et al., 1991] Neches, R., Fikes, R., Finin, T., Gruber, T., Patil, R., Senator, T., and Swartout, W. R. (1991). Enabling technology for knowledge sharing. *AI Mag.*, 12(3) :36–56. 21

[Neveu and Quéré, 1996] Neveu, E. and Quéré, L. (1996). *Le temps de l'événement I*, chapter Présentation, pages 7–21. Number 75 in Réseaux. CNET. 27, 75

[Newcombe et al., 1959] Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. (1959). Automatic Linkage of Vital Records. *Science*, 130(3381) :954–959. 63

[Nishihara et al., 2009] Nishihara, Y., Sato, K., and Sunayama, W. (2009). Event extraction and visualization for obtaining personal experiences from blogs. In *Proceedings of the Symposium on Human Interface 2009 on Human Interface and the Management of Information. Information and Interaction. Part II : Held as part of HCI International 2009*, pages 315–324, Berlin, Heidelberg. Springer-Verlag. 54

[NIST, 2005] NIST (2005). *The ACE 2005 (ACE05) Evaluation Plan*. 28

[Noël, 2008] Noël, L. (2008). From semantic web data to inform-action : a means to an end. Workshop SWUI (Semantic Web User Interaction), CHI 2008, 5-10 Avril, 2008, Florence, Italie. 136

[Noy and Mcguinness, 2001] Noy, N. F. and Mcguinness, D. L. (2001). Ontology Development 101 : A Guide to Creating Your First Ontology. Technical Report KSL-01-05, Stanford Knowledge Systems Laboratory. 22, 78, 79

[Padró and Stanilovsky, 2012] Padró, L. and Stanilovsky, E. (2012). Freeling 3.0 : Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey. ELRA. 51

[Paquet, 2008] Paquet, P. (2008). De l'information à la connaissance. In *Information et communication et management dans l'entreprise : quels enjeux ?*, pages 17–48. Harmattan. 18

[Paumier, 2003] Paumier, S. (2003). A Time-Efficient Token Representation for Parsers. In *Proceedings of the EACL Workshop on Finite-State Methods in Natural Language Processing*, pages 83–90, Budapest. 52

[Pauna and Guillemin-Lanne, 2010] Pauna, R. and Guillemin-Lanne, S. (2010). Comment le text mining peut-il aider à gérer le risque militaire et stratégique ? *Text*. 53

[Piskorski and Atkinson, 2011] Piskorski, J. and Atkinson, M. (2011). Frontex real-time news event extraction framework. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 749–752, New York, NY, USA. ACM. 53

[Piskorski et al., 2007] Piskorski, J., Tanev, H., and Wennerberg, P. O. (2007). Extracting violent events from on-line news for ontology population. In *Proceedings of the 10th international conference on Business information systems*, BIS'07, pages 287–300, Berlin, Heidelberg. Springer-Verlag. 54

[Piskorski and Yangarber, 2013] Piskorski, J. and Yangarber, R. (2013). Information extraction : Past, present and future. In Poibeau, T., Saggion, H., Piskorski, J., and Yangarber, R., editors, *Multi-source, Multilingual Information Extraction and Summarization*, Theory and Applications of Natural Language Processing, pages 23–49. Springer Berlin Heidelberg. 57

[Poibeau, 2003] Poibeau, T. (2003). *Extraction automatique d'information : Du texte brut au web sémantique*. Lavoisier. 40

[Polanyi, 1966] Polanyi, M. (1966). *The tacit dimension*. Routledge and Keagan Paul. 18

[Popescu et al., 2011] Popescu, A.-M., Pennacchiotti, M., and Paranjpe, D. (2011). Extracting events and event descriptions from Twitter. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11, pages 105–106, New York, NY, USA. ACM. 54

[Pustejovsky et al., 2003] Pustejovsky, J., Castaño, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A., Katz, G., and R., R. D. (2003). TimeML : Robust specification of event and temporal expressions in text. In *New Directions in Question Answering'03*, pages 28–34. 27

[Quine, 1985] Quine, W. V. (1985). Events and reification. In *Actions and Events : Perspectives on the Philosophy of Davidson*, pages 162–71. Blackwell. 66

[Quine, 1960] Quine, W. V. O. (1960). *Word and Object*. MIT Press paperback series. Technology Press of the Massachusetts Inst. of Technology. 26

[Radev et al., 2001] Radev, D. R., Blair-Goldensohn, S., Zhang, Z., and Raghavan, R. S. (2001). Interactive, domain-independent identification and summarization of topically related news articles. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '01, pages 225–238, London, UK, UK. Springer-Verlag. 54

[Raimond et al., 2007] Raimond, Y., Abdallah, S. A., Sandler, M. B., and Giasson, F. (2007). The music ontology. In *ISMIR*, pages 417–422. 28

[Randell et al., 1992] Randell, D. A., Cui, Z., and Cohn, A. G. (1992). A spatial logic based on regions and connection. In *Proceedings of the 3rd international conference on knowledge representation and reasoning*. 35

[Ratinov et al., 2011] Ratinov, L., Roth, D., Downey, D., and Anderson, M. (2011). Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies - Volume 1*, HLT '11, pages 1375–1384, Stroudsburg, PA, USA. Association for Computational Linguistics. 58

[Rattenbury et al., 2007] Rattenbury, T., Good, N., and Naaman, M. (2007). Towards automatic extraction of event and place semantics from flickr tags. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 103–110, New York, NY, USA. ACM. 54

[Ricœur, 1983] Ricœur, P. (1983). *Temps et récit 1. L´intrigue et le récit historique*, volume 227 of *Points : Essais*. Ed. du Seuil, Paris. 27

[Rosario and Hearst, 2004] Rosario, B. and Hearst, M. A. (2004). Classifying semantic relations in bioscience texts. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*, pages 430–es. 47

[Rosario and Hearst, 2005] Rosario, B. and Hearst, M. A. (2005). Multi-way relation classification : application to protein-protein interactions. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 732–739, Stroudsburg, PA, USA. Association for Computational Linguistics. 41, 53

[Saurí et al., 2005] Saurí, R., Knippen, R., Verhagen, M., and Pustejovsky, J. (2005). Evita : a robust event recognizer for QA systems. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 700–707, Stroudsburg, PA, USA. Association for Computational Linguistics. 49, 54

[Saval, 2011] Saval, A. (2011). *Modèle temporel, spatial et sémantique pour la découverte de relations entre évènements*. PhD thesis, Univeristé de Caen Basse-Normandie. 27

[Saval et al., 2009] Saval, A., Bouzid, M., and Brunessaux, S. (2009). A Semantic Extension for Event Modelisation. *2009 21st IEEE International Conference on Tools with Artificial Intelligence*, pages 139–146. 74, 75

[Sayyadi et al., 2009] Sayyadi, H., Hurst, M., and Maykov, A. (2009). Event Detection and Tracking in Social Streams. In *Proceedings of International Conference on Weblogs and Social Media (ICWSM)*. 54

[Saïs et al., 2009] Saïs, F., Pernelle, N., and Rousset, M.-C. (2009). Combining a logical and a numerical method for data reconciliation. In Spaccapietra, S., editor, *Journal on Data Semantics XII*, volume 5480 of *Lecture Notes in Computer Science*, pages 66–94. Springer Berlin Heidelberg. 63

[Schmid, 1994] 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. 99

[Sekine et al., 2002] Sekine, S., Sudo, K., and Nobata, C. (2002). Extended named entity hierarchy. In Rodríguez, M. G. and Araujo, C. P. S., editors, *Proceedings of* 3^{rd} *International Conference on Language Resources and Evaluation (LREC'02)*, pages 1818–1824, Canary Islands, Spain. 58

[Serrano et al., 2013a] Serrano, L., Bouzid, M., Charnois, T., Brunessaux, S., and Grilheres, B. (2013a). Events extraction and aggregation for open source intelligence : from text to knowledge. In *IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2013)*, Washington DC, USA.

[Serrano et al., 2013b] Serrano, L., Bouzid, M., Charnois, T., Brunessaux, S., and Grilheres, B. (2013b). Extraction et agrégation automatique d'événements pour la veille en sources ouvertes : du texte à la connaissance. In *Ingénierie des Connaissances 2013 (IC 2013)*, Lille, France.

[Serrano et al., 2012a] Serrano, L., Bouzid, M., Charnois, T., and Grilheres, B. (2012a). Vers un système de capitalisation des connaissances : extraction d'événements par combinaison de plusieurs approches. In *Atelier des Sources Ouvertes au Web de Données (SOS-DLWD'2012) en conjonction avec la conférence internationale francophone (EGC 2012)*, Bordeaux, France.

[Serrano et al., 2012b] Serrano, L., Charnois, T., Brunessaux, S., Grilheres, B., and Bouzid, M. (2012b). Combinaison d'approches pour l'extraction automatique d'événements (automatic events extraction by combining multiple approaches) [in french]. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2 : TALN*, Grenoble, France. ATALA/AFCP.

[Serrano et al., 2011] Serrano, L., Grilheres, B., Bouzid, M., and Charnois, T. (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)*, Brest, France.

[Silberztein et al., 2012] Silberztein, M., Váradi, T., and Tadic, M. (2012). Open source multi-platform nooj for nlp. In *COLING (Demos)*, pages 401–408. 52

[Smart et al., 2007] Smart, P., Russell, A., Shadbolt, N., Shraefel, M., and Carr, L. (2007). Aktivesa. *Comput. J.*, 50 :703–716. 36

[Soderland, 1999] Soderland, S. (1999). Learning information extraction rules for semi-structured and free text. *Mach. Learn.*, 34 :233–272. 43

[Sun et al., 2005] Sun, Z., Lim, E.-P., Chang, K., Ong, T.-K., and Gunaratna, R. K. (2005). Event-driven document selection for terrorism information extraction. In *Proceedings of the 2005 IEEE international conference on Intelligence and Security Informatics*, ISI'05, pages 37–48, Berlin, Heidelberg. Springer-Verlag. 53

[Tanev et al., 2008] Tanev, H., Piskorski, J., and Atkinson, M. (2008). Real-time news event extraction for global crisis monitoring. In *Proceedings of the 13th international conference on Natural Language and Information Systems : Applications of Natural Language to Information Systems*, NLDB '08, pages 207–218, Berlin, Heidelberg. Springer-Verlag. 53

[Tarski, 1956] Tarski, A. (1956). *Logic, Semantics, Metamathematics*, chapter Foundations of the Geometry of Solids. Oxford, Clarendon Press. 35

[Thompson et al., 1999] Thompson, C. A., Califf, M. E., and Mooney, R. J. (1999). Active learning for natural language parsing and information extraction. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pages 406–414, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. 58

[Tkachenko and Simanovsky, 2012] Tkachenko, M. and Simanovsky, A. (2012). Named entity recognition : Exploring features. In Jancsary, J., editor, *Proceedings of KONVENS 2012*, pages 118–127. ÖGAI. Main track : oral presentations. 45

[Troncy et al., 2010] Troncy, R., Shaw, R., and Hardman, L. (2010). LODE : une ontologie pour représenter des événements dans le web de données. In *IC 2010, 21st Journées Francophones d'Ingénierie des Connaissances, June 8-11, 2010, Nîmes, France*, Nîmes, FRANCE. 29

[Van De Velde, 2006] Van De Velde, D. (2006). *Grammaire des événements*. Presses Universitaires du Septentrion. 26

[van Hage et al., 2011] van Hage, W. R., Malaisé, V., Segers, R., Hollink, L., and Schreiber, G. (2011). Design and use of the Simple Event Model (SEM). *Web Semantics : Science, Services and Agents on the World Wide Web*, 9(2) :128–136. 30

[Vargas-Vera and Celjuska, 2004] Vargas-Vera, M. and Celjuska, D. (2004). Event recognition on news stories and semi-automatic population of an ontology. In *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '04, pages 615–618, Washington, DC, USA. IEEE Computer Society. 54

[Varjola and Löffler, 2010] Varjola, M. and Löffler, J. (2010). PRONTO : Event Recognition for Public Transport. *Proceedings of 17th ITS World Congress, Busan, Korea*. 53

[Verykios and Elmagarmid, 1999] Verykios, V. S. and Elmagarmid, A. K. (1999). Automating the approximate record matching process. *Information Sciences*, 126 :83–98. 64

[Viola and Narasimhan, 2005] Viola, P. and Narasimhan, M. (2005). Learning to extract information from semi-structured text using a discriminative context free grammar. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 330–337, New York, NY, USA. ACM. 45

[Vlahovic, 2011] Vlahovic, N. (2011). Information Retrieval and Information Extraction in Web 2.0 environment. *nternational Journal of Computers*, 5(1). 40, 54

[Wacholder et al., 1997] Wacholder, N., Ravin, Y., and Choi, M. (1997). Disambiguation of proper names in text. In *Proceedings of the fifth conference on Applied natural language processing*, ANLC '97, pages 202–208, Stroudsburg, PA, USA. Association for Computational Linguistics. 64

[Wakao et al., 1996] Wakao, T., Gaizauskas, R., and Wilks, Y. (1996). Evaluation of an algorithm for the recognition and classification of proper names. In *Proceedings of the 16th conference on Computational linguistics - Volume 1*, COLING '96, pages 418–423, Stroudsburg, PA, USA. Association for Computational Linguistics. 44

[Wang et al., 2011] Wang, W., Besançon, R., Ferret, O., and Grau, B. (2011). Filtering and clustering relations for unsupervised information extraction in open domain. *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*, page 1405. 47

[Whitehead, 1920] Whitehead, A. (1920). *The Concept of Nature*. Dover science books. Dover Publications. 35

[Widlocher et al., 2006] Widlocher, A., Bilhaut, F., Hernandez, N., Rioult, F., Charnois, T., Ferrari, S., and Enjalbert, P. (2006). Une approche hybride de la segmentation thématique : collaboration du traitement automatique des langues et de la fouille de texte. In *Actes de DEfi Fouille de Texte (DEFT'06), Semaine du Document Numérique (SDN'06)*, Fribourg, Suisse. 52

[Winkler et al., 2006] Winkler, W. E., Winkler, W. E., and P, N. (2006). Overview of record linkage and current research directions. Technical report, Bureau of the Census. 63, 109

[Xu et al., 2006] Xu, F., Uszkoriet, H., and Li, H. (2006). Automatic Event and Relation Detection with Seeds of Varying Complexity. In *AAAI 2006 Workshop on Event Extraction and Synthesis*. 49

[Yates and Etzioni, 2009] Yates, A. and Etzioni, O. (2009). Unsupervised methods for determining object and relation synonyms on the web. *J. Artif. Intell. Res. (JAIR)*, 34 :255–296. 66

[Zanasi, 2009] Zanasi, A. (2009). Virtual weapons for real wars : Text mining for national security. In Corchado, E., Zunino, R., Gastaldo, P., and Herrero, A., editors, *Proceedings of the International Workshop on Computational Intelligence in Security for Information Systems CISIS'08*, volume 53 of *Advances in Soft Computing*, pages 53–60. Springer Berlin Heidelberg. 53

[Zhao, 2004] Zhao, S. (2004). Named entity recognition in biomedical texts using an hmm model. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, JNLPBA '04, pages 84–87, Stroudsburg, PA, USA. Association for Computational Linguistics. 45

[Zhu and Porter, 2002] Zhu, D. and Porter, A. L. (2002). Automated extraction and visualization of information for technological intelligence and forecasting. *Technological Forecasting and Social Change*, 69(5) :495 – 506. <ce :title>TF Highlights from {ISF} 2001</ce :title>. 53

[Zhu et al., 2009] Zhu, J., Nie, Z., Liu, X., Zhang, B., and Wen, J.-R. (2009). StatSnowball : a statistical approach to extracting entity relationships. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 101–110, New York, NY, USA. ACM. 47