ANNOTATION D'EVENEMENTS DANS LES TEXTES POUR LA VEILLE STRATEGIQUE

Bénédicte GOUJON
benedicte.goujon@thalesgroup.com

Thales Research & Technology France
Campus de Polytechnique – 1, avenue Augustin Fresnel
91 767 Palaiseau Cedex

Mots clefs :
Veille stratégique ; annotation textuelle ; analyse linguistique ; services web

Keywords:
Strategic intelligence; textual annotation; linguistic analysis; web services

Résumé

L'objectif de ce travail est de proposer l'annotation automatique d'événements dans des textes, en identifiant la certitude associée à ces événements, pour l'aide à la veille stratégique. Notre sujet de travail est la crise de septembre 2002 en Côte d'Ivoire. A partir de l'analyse automatique des dépêches de presse précédant cet événement, nous souhaitons identifier de façon précise les événements crisogènes (événements susceptibles de provoquer ou de révéler une crise) qui ont eu lieu, dans l'objectif d'aider à une éventuelle anticipation de ces situations. L'identification des événements crisogènes a consisté à lister dans un premier temps un ensemble d'événements crisogènes, à capitaliser leurs formes textuelles dans un deuxième temps grâce à notre outil SemPlusEvent de génération de patrons linguistiques, puis à identifier avec quelle certitude ou quelle réalité cet événement est annoncé. Ce travail se situe dans le cadre du projet ANR WebContent, dont l'objectif est de mettre en œuvre une plateforme réunissant des services web d'analyse des informations textuelles pour des applications de veille. Une implémentation de la méthode présentée a été réalisée sous la forme de services web.

1 Introduction

L'objectif de ce travail est de proposer l'annotation automatique d'événements dans des textes, en identifiant la certitude associée à ces événements, pour l'aide à la veille stratégique. Notre sujet de travail est la crise de septembre 2002 en Côte d'Ivoire. A partir de l'analyse automatique des dépêches de presse précédant cet événement, nous souhaitons identifier de façon précise les événements crisogènes (événements susceptibles de provoquer ou de révéler une crise) qui ont eu lieu, dans l'objectif d'aider à une éventuelle anticipation de ces situations. L'identification précise des événements crisogènes a consisté à lister dans un premier temps un ensemble d'événements crisogènes, à capitaliser leurs formes textuelles dans un deuxième temps, puis à identifier avec quelle certitude ou quelle réalité cet événement est annoncé. Notre approche s'appuie sur l'analyse linguistique des textes pour l'extraction automatique d'événements.

Ce travail se situe dans le cadre du projet ANR WebContent[1], dont l'objectif est de mettre en œuvre une plateforme réunissant des services web d'analyse des informations textuelles pour des applications de veille.

Nous présentons tout d'abord notre problématique de veille stratégique que nous devons traiter, sur la Côte d'Ivoire. Ensuite, nous détaillons différents aspects de notre méthode d'annotation des textes :

- l'identification des événements crisogènes (meurtre, rencontre, …) ;
- l'acquisition des formes textuelles associées aux entités nommées qui interviennent dans les événements (« Le président », « Laurent Gbagbo », …) ;
- la capitalisation des formes textuelles de chaque événement avec SemPlus, notre outil d'acquisition de patrons linguistiques par apprentissage ;
- le repérage de l'incertitude exprimée par l'auteur du texte.

Puis nous présentons l'implémentation effectuée dans le cadre de la plate-forme WebContent, sous la forme de services web, et l'utilisation des résultats produits dans des applications de veille. La problématique de la validation de notre approche est enfin présentée.

2 La problématique de veille stratégique

Notre problématique est la suivante : comment utiliser au mieux les sources d'informations textuelles pour aider un expert en veille stratégique à connaître le contexte qui l'intéresse ? Comment l'aider à capturer toutes les nuances apportées par les auteurs des textes sans l'obliger à lire tous les textes ? La masse d'informations disponibles aujourd'hui, même sur un sujet bien défini (exemple : la situation politique actuelle en Côte d'Ivoire), est très vaste et généralement les experts n'ont pas les moyens de tout lire. En même temps, ils ne peuvent risquer de passer à côté de l'information importante. Pour aider ces experts, les systèmes d'analyse automatique des textes sont une solution aujourd'hui indispensable.

Dans notre travail, nous proposons l'annotation automatique des textes en fonction d'événements crisogènes qui intéressent a priori l'expert. Nous travaillons sur le cas particulier de la crise en Côte d'Ivoire qui a eu lieu en septembre 2002, avec le 19 septembre une tentative de coup d'état qui a dégénéré en soulèvement armé. Notre objectif est d'aider l'expert à connaître le contexte géo-politique du moment grâce à une extraction automatique des événements crisogènes à partir des dépêches concernant la période précédant la crise. Nous travaillons sur un corpus en français, mais l'approche présentée ici est aussi valable sur l'anglais.

[1] [1] WebContent, www.webcontent.fr/.

3 La méthode d'annotation sémantique des textes

Le premier résultat que nous visons est l'annotation sémantique des événements crisogènes décrits dans les textes. Cette annotation se fait en plusieurs étapes :

- Des étapes de préparation des connaissances ontologiques et linguistiques :
  - L'identification des événements crisogènes ;
  - L'acquisition des formes textuelles associées aux participants d'événements (entités nommées telles que personnes, lieux, organisation, …) ;
  - L'acquisition des formes textuelles associées à chaque événement crisogène ;
  - La caractérisation de l'incertitude exprimée.
- Des étapes d'utilisation de ces connaissances :
  - L'annotation des entités nommées, participants aux événements ;
  - L'annotation des événements avec l'incertitude associée.

Nous présentons ci-après chacun de ces points.

3.1 Identification des événements crisogènes

Une première étape consiste à identifier les événements crisogènes. Des classifications des événements crisogènes ont déjà été définies. Par exemple, le tableau suivant reprend la typologie simplifiée des troubles de l'ordre public proposée par le gouvernement tchadien[2].

[2] Cité dans le livrable Infom@gic : ST3.41 : Application en gestion des risques. Premiers modèles d'évaluation du risque appliqués à la montée de la tension socio-politique en Côte d'Ivoire.

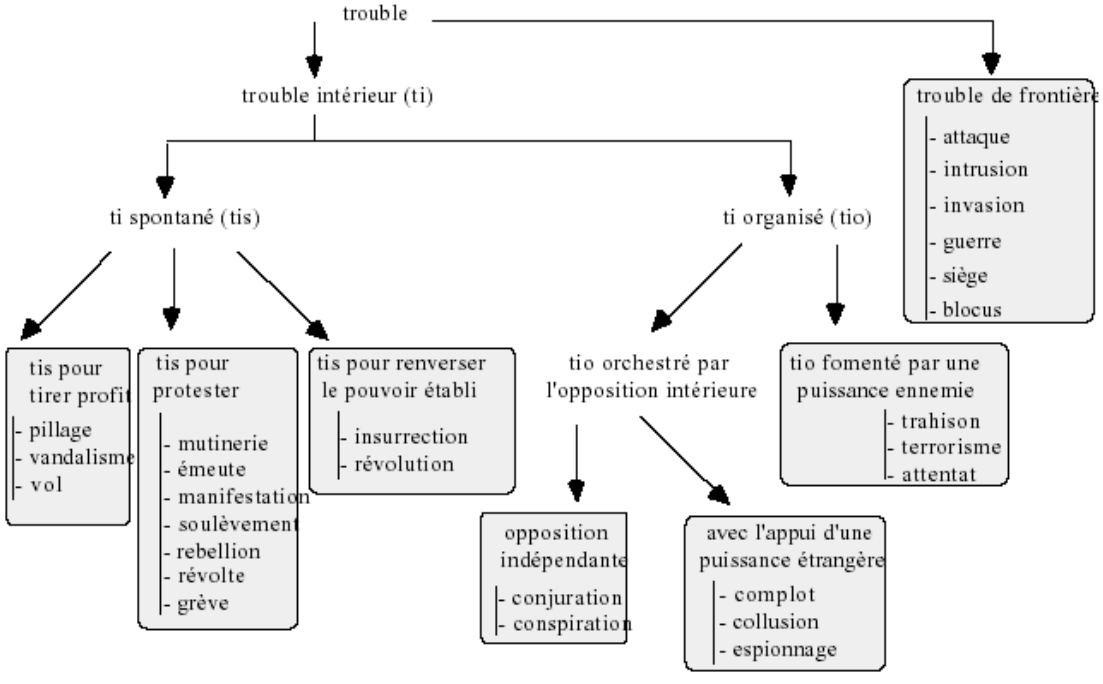

Figure 1 : Une classification d'événements crisogènes par le gouvernement tchadien.

Cette classification très intéressante n'a pas été utilisée directement dans notre approche. En effet, cette classification ne tient par exemple pas compte d'événements plus anodins tels que des rencontres ou des déplacements qui peuvent dans certains cas entraîner des crises (rencontre entre deux opposants, déplacement d'un groupe armé vers un lieu stratégique, …). De même, certains événements à l'initiative de représentants du pouvoir (descente de police) peuvent entraîner des crises. Dans notre travail, nous nous sommes centrés sur un sous-ensemble d'événements, mais une évaluation à grande échelle pourrait s'appuyer sur la prise en compte de l'ensemble des catégories présentées ci-dessus plus quelques catégories supplémentaires. Pour notre étude de cas, nous avons étudié les dépêches et articles sur la Côte d'Ivoire datés entre le 1er et le 18 septembre 2002, soit 203 documents. A partir de ces documents, nous avons établi une première liste d'événements potentiellement crisogènes :

- Assassinat
- Vol
- Rencontre
- Déplacement
- Descente de police

- Agression
- Combat
- Arrestation

3.2 Acquisition des formes textuelles associées aux entités nommées

Le repérage d'événements dans les textes s'appuie sur une première étape qui consiste à repérer les participants de ces événements. Ces participants sont principalement désignés par des entités nommées : lieux, personnes, organisations, dates, … L'acquisition des formes textuelles associées à ces entités peut se faire de deux façons : soit par saisie manuelle, soit par apprentissage automatique. L'apprentissage de formes textuelles associées à des personnes ou des organisations a par exemple été présenté dans la thèse de T. Poibeau en 2002 [1]. Il s'appuie sur le repérage d'amorces (M., Monsieur, la société) qui, suivies de mots avec majuscule, introduisent des entités nommées. D'autres approches se limitent au repérage de noms propres (mots avec majuscule), sans chercher à les typer automatiquement. Aujourd'hui, ces méthodes se sont répandues et sont présentes dans des produits du commerce.
De notre point de vue, la difficulté de cette tâche consiste à repérer toutes les formes textuelles associées à chaque entité nommée. Par exemple, si l'on repère « Le président a rencontré Alassane Ouattara », il est nécessaire de savoir de quel président il est question pour repérer cette relation et ses participants. La résolution des anaphores sur les personnes fait actuellement l'objet d'une thèse à Thales [2]. L'objectif est non seulement de résoudre les anaphores pronominales telles que « il », les anaphores nominales telles que « le président », mais aussi de trouver les référents pour des groupes de personnes tels que « les trois hommes ». Nous travaillons aussi sur le repérage des référents calendaires des expressions temporelles telles que « hier », « lundi dernier » en s'appuyant sur la date du document, qui vise un objectif similaire : permettre la compréhension hors contexte d'une information extraite d'un texte.

3.3 Capitalisation des formes textuelles d'événements avec SemPlus

Cette étape consiste à lister les formes textuelles que peut prendre chaque événement crisogène. Par exemple, un « meurtre » peut être exprimé par « assassinat de X », « X a été tué par Y », « Y a froidement abattu X … », « Le meurtre de X … ». L'acquisition de ces formes est facilitée par l'utilisation de notre outil SemPlusEvent, qui a pour but de permettre à un non-linguiste de créer des patrons syntaxico-sémantiques associés à des événements à partir d'exemples.
La version initiale de cet outil, nommé SemPlus, permet l'acquisition de patrons d'extractions associés à des relations binaires, telles que « X rencontre Y » pour la relation de Rencontre ou « X s'est rendu en Y » pour la relation de Déplacement. SemPlus, qui a été mis au point lors d'une collaboration avec la STAT (Services Techniques de l'Armée de Terre), a été réalisé afin de permettre à des non-linguistes (experts en veille stratégiques) de pouvoir saisir sans difficultés, de façon autonome, des patrons linguistiques [3], [4]. SemPlus contient deux étapes : une étape d'apprentissage des patrons, une étape d'application des patrons pour l'extraction de relations. Pour l'apprentissage, qui est un apprentissage symbolique, l'outil s'appuie sur un parcours d'un premier ensemble de documents appelé corpus d'apprentissage.

SemPlus met en œuvre l'algorithme d'apprentissage suivant, proche de celui de Hearst [5] :

1. Choix par l'utilisateur du couple de catégories d'entités en jeux dans la ou les relations pertinentes. Ainsi, seules les phrases contenant des instances des catégories choisies sont présentées à l'utilisateur.

2. Saisie par l'utilisateur de couples d'instances vérifiant la relation recherchée.

3. Récupération automatique des phrases du corpus d'apprentissage contenant ces couples, donc avec des patrons décrivant potentiellement les relations recherchées.

4. Copie par l'utilisateur des extraits de phrases exprimant les relations, transformation automatique de ces extraits en patrons linguistiques. Les patrons sont des graphes au format de l'environnement linguistique Intex [6]. A cette étape, l'utilisateur peut modifier l'extrait, en remplaçant des suite de mots non pertinentes par « ** ».

5. Application automatique des patrons sur le corpus d'apprentissage : récupération automatique de nouveaux couples. Retour à l'étape 2.

Pour illustrer l'approche de SemPlus, voici un exemple. L'utilisateur s'intéresse aux relations de type rencontre ou contact entre deux personnes. Après la réduction du corpus d'apprentissage (1.) pour ne conserver que les phrases contenant au moins deux personnes, l'utilisateur repère la phrase suivante « … Jacques Chirac a téléphoné mercredi à Laurent Gbagbo … ». L'utilisateur saisit alors le couple « Jacques Chirac » - « Laurent Gbagbo » (2.) et récupère toutes les phrases contenant ces deux personnes (3.). En reprenant la phrase précédemment repérée, l'utilisateur produit l'extrait suivant : « Jacques Chirac a téléphoné ** à Laurent Gbagbo » (4.), qui est transformé automatiquement en graphe Intex pouvant s'exprimer sous la forme : « Personne1 <téléphoner> ** à Personne2 => Contact(Agent : Personne1, Patient : Personne2) ». L'utilisation de ce patron sur un autre corpus contenant « Jacques Chirac aurait téléphoné le 17 juillet à Vojislav Kostunica … » entraîne l'identification de la nouvelle relation : Contact(Agent : Jacques Chirac, Patient : Vojislav Kostunica). A partir de ce nouveau couple, le système peut mettre en valeur une nouvelle phrase exprimant la même relation : « Jacques Chirac a dit à Paris qu'il avait invité Vojislav Kostunica … » (étape 3), qui permettra la capture d'un nouveau patron : « Personne1 ** <inviter> Personne2 » => Contact(Agent : Personne1, Patient : Personne 2) ».
La nouvelle version de cet outil, nommée SemPlusEvent [7], permet de traiter des événements n-aires, par exemple : Contact(Agent, Patient, Date, Lieu) ou Meutre(Patient, Date, Lieu), où la date et le lieu sont des informations optionnelles, qui ne sont pas toujours indiquées. L'algorithme et l'interface ont été modifié afin de prendre en compte ces améliorations, mais le but reste de faciliter l'acquisition de connaissances linguistiques par des non-linguistes. La figure ci-après montre l'interface d'affichage des événements extraits automatiquement de SemPlusEvent.

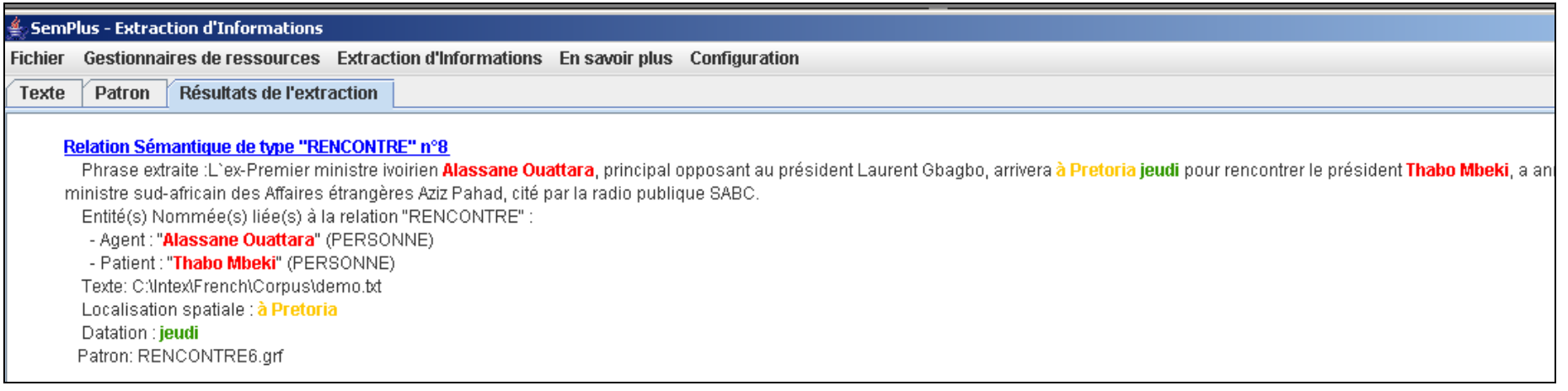

Figure 2 : Interface de SemPlusEvent montrant un événement extrait.

3.4 Caractérisation de l'incertitude

L'extraction automatique d'événements à partir de textes doit tenir compte du contexte discursif dans lequel ces événements sont présentées. Ainsi, « Selon un témoin, Y a abattu X », « Y a peut-être tué X », « X a été tué par Y » expriment trois degrés différents de certitude de l'auteur principal du texte. Dans le premier cas, l'auteur présente un discours rapporté, qui exprime une certaine distance, incertitude de l'auteur principal vis-à-vis de l'événement. Ici, la certitude de l'événement est relative à la fiabilité de la source secondaire (« un témoin » dans cet exemple). Dans le deuxième cas, l'auteur exprime une incertitude via « peut-être ». Dans le troisième cas, l'auteur n'exprime aucune incertitude. Ces nuances, évidentes lors de la lecture des textes, doivent être prises en compte par les systèmes d'extraction automatique d'événements. Les premiers besoins en extraction d'événements, exprimés via les campagnes MUC 3 (1991) et MUC 4 (1992) du NIST visaient l'identification des lieux, dates, auteurs et victimes d'attentats passés [1], donc peu liés à des expressions d'incertitude. Plus récemment, les campagnes ACE (Automatic Content Extraction) ont intégré l'extraction d'événements, avec le repérage d'attributs (modalité, polarité). Mais, en 2007 [8], un seul candidat (BBN Technologies) a tenté cette épreuve, qui a été supprimé en 2008. On peut en déduire que les systèmes ne sont pas encore prêts pour de telles évaluations.

L'importance de la prise en compte de l'incertitude exprimée par l'auteur du texte est aujourd'hui bien identifiée, ainsi Auger et Roy de la Defense R&D Canada [9] montrent que pour permettre à l'étape suivante la fusion d'informations, il est nécessaire en amont de prendre en compte les ambiguïtés ainsi que les expressions de certitude/incertitude exprimées dans les textes.

Notre approche s'appuie sur l'exploitation de l'ensemble des marqueurs linguistiques et des structures linguistiques permettant d'exprimer la réalisation, ainsi que les marqueurs et structures linguistiques associés à l'incertitude et au discours rapporté. Voici des marqueurs qui expriment la non réalisation : temps futur ou négation, verbes modaux qui introduisent un événement (empêcher que, souhaiter que). Pour exprimer l'incertitude quant à la réalisation ou non d'un événement, plusieurs marqueurs sont utilisés : temps conditionnel, adverbes (peut-être), Si …, et les discours rapportés. La spécificité des discours rapporté est qu'une source est parfois indiquée, ce qui modifie fortement la valeur de certitude que l'utilisateur peut associer à une information, tandis que la valeur d'un conditionnel non associée à une source sera toujours la même.

Quatre caractéristiques sont attribuées aux segments textuels. Parmi ces quatre caractéristiques, trois ont un ensemble fini de valeurs, qui peuvent être adaptés selon les applications visées :

- Temps : passé, en cours, futur.
- Incertitude : grande incertitude, incertitude moyenne, faible incertitude.
- Polarité : affirmation, négation.
- Source : contenu correspondant à la source locale

Par défaut, la valeur de Temps est « passé », et la valeur de Polarité est « affirmation ».
Dans les travaux existants, Rubin [10] propose un modèle de l'incertain qui est proche de notre modèle, mais qui ne tient par exemple pas compte de la négation.
Les exemples suivants montrent les résultats visés en terme de caractérisation de l'incertitude exprimée dans les phrases.

- « Laurent Gbagbo devrait se rendre en Italie. » => Temps : futur, Incertitude : incertitude moyenne.
- « Laurent Gbagbo ne se rendra sûrement pas la semaine prochaine en Italie. » => Temps : futur, Incertitude : forte incertitude.
- « Selon le leader des rebelles, Laurent Gbagbo s'est rendu hier en Italie. » => Temps : passé, Incertitude : incertitude moyenne, source : le leader des rebelles.
- « Laurent Gbagbo ne s'est pas rendu en Italie. » => Temps : passé, Polarité : négation.

Dans le deuxième exemple ci-dessus, la combinaison de la négation et de la forte certitude est analysée pour produire la valeur de forte incertitude. Sur le dernier exemple, la polarité va permettre de ne pas tenir compte d'événements présentés comme n'ayant pas eu lieu.

4 Implémentation et exploitation des résultats

4.1 Ressources

Notre programme nommé SemPlusEEWS (SemPlus Event Extraction Web Service), développé sous la forme d'un service web, prend en entrée une ontologie. Le schéma suivant montre l'organisation de notre ontologie.

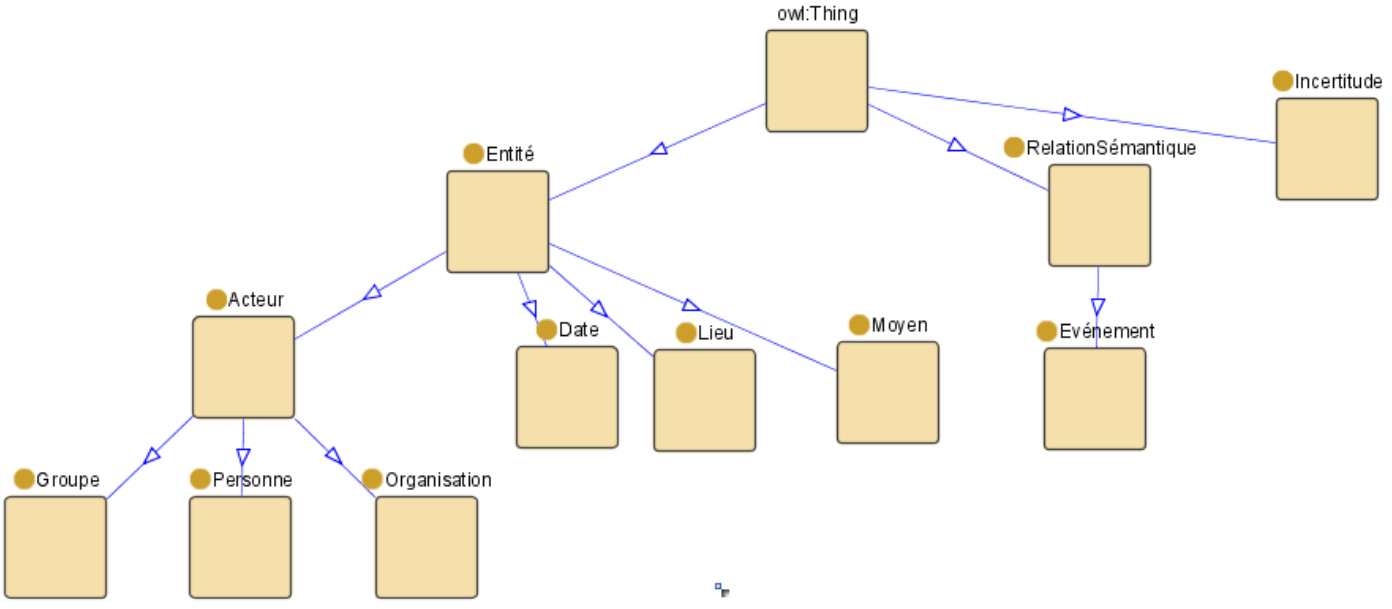

Figure 3 : Schéma de l'ontologie utilisée par nos services web.

Sur le cas de la Côte d'Ivoire qui nous intéresse, l'ontologie contient 118 personnes, 50 organisations et 89 lieux. Ces informations ne sont pas uniquement liées à la Côte d'Ivoire, puisque sur la période analysé le président ivoirien s'est rendu en Italie, où il a rencontré plusieurs personnalités politiques. Cette ontologie, qui contient de nombreuses instances de personnes, groupes ou lieux, est transformée en dictionnaires au format Dela pour une exploitation par Intex lors de l'analyse de textes.

```
Emile Boga Doudou,.PERSONNE
Félix Houphouët Boigny,.PERSONNE
Gbagbo,Laurent Gbagbo.PERSONNE
Henri Konan Bédié,.PERSONNE
Henriette Dagri Diabaté,.PERSONNE
Henry Aussavy,colonel Henry Aussavy.PERSONNE
Houphouët-Boigny,Félix Houphouët Boigny.PERSONNE
IB,Ibrahim Coulibaly.PERSONNE
```

Figure 4 : Extrait du dictionnaire de personnes généré à partir de l'ontologie.

Dans ce dictionnaire, des variantes peuvent être associées à une forme générique. La forme générique est alors en deuxième position (par exemple, Laurent Gbagbo est la forme générique, Gbagbo est une variante). Le dictionnaire peut contenir plusieurs variantes textuelles pour désigner une même instance.

4.2 Annotation des événements dans les textes

Après avoir configuré notre service avec l'ontologie de référence, SemPlusEEWS est appelé avec le texte à analyser, qui est encapsulé dans un élément (MediaUnit) spécifique au format WebContent (descriptions XML). Pour l'annotation, SemPlusEEWS appelle les patrons linguistiques construits préalablement avec notre outil SemPlusEvent.
Les annotations produites par notre service SemPlusEEWS sont sous la forme RDF. RDF est un format qui s'appuie sur une description en triplet des informations, avec un balisage du type XML. C'est un langage du web sémantique, dont l'objectif est de décrire l'information présente dans les textes de façon structurée afin qu'elle soit accessible pour des machines.

Pour illustrer nos résultats, nous avons analysé le texte fictif suivant, concentrant deux événements :

Il y a eu hier un hold-up à la BCEAO. L'assassinat de Balla Kéïta s'est produit le 13 juillet 2002 à Ouagadougou selon nos sources.

L'analyse de ce court texte a produit les annotations RDF suivantes (en gras, les informations importantes) :

```
...
<rdf:Description rdf:about="weblab://InstanceCandidate//Incertitude#0">
<onto:valeur_d_incertitude>incertitude_moyenne</onto:valeur_d_incertitude>
<onto:source>nos sources</onto:source>
</rdf:Description>
</rdf:RDF>
...
<rdf:Description rdf:about="weblab://InstanceCandidate//Evenement#0">
<rdf:type rdf:resource="ASSASSINAT"/>
<onto:Patient>http://www.owl-ontologies.com/RCI.owl#Personne_16</onto:Patient>
<onto:localisation>à Ouagadougou</onto:localisation>
<onto:datation>le 13 juillet 2002</onto:datation>
<onto:incertitude_liee>weblab://InstanceCandidate//Incertitude#0</onto:incertitude_liee>
</rdf:Description>
</rdf:RDF>
...
<rdf:Description rdf:about="weblab://InstanceCandidate//Evenement#1">
<rdf:type rdf:resource="HOLD_UP"/>
<onto:Patient>http://www.owl-ontologies.com/RCI.owl#Organisation3</onto:Patient>
<onto:datation>hier</onto:datation>
</rdf:Description>
</rdf:RDF>
...
```

Voici ce à quoi correspond la première partie de ces annotations : l'événement Evenement#0 est de type ASSASSINAT. Le Patient de cet événement est Personne_16, qui est une instance de l'ontologie de la Côte d'Ivoire (RCI.owl) qui correspond à Balla Kéïta. La localisation de cet événement est « à Ouagadougou », la datation est « le 13 juillet 2002 », et l'incertitude liée (Incertitude#0) a une valeur d'incertitude « incertitude moyenne », avec en source « nos sources ». D'autres annotations du document, non montrées ici, permettent de relier chacune de ces informations aux segments textuels sources.
Dans la version actuelle de notre service d'annotation des entités nommées, la localisation et la datation des événements sont assez sommaires : la localisation ne fait pas référence aux instances de lieux définies dans l'ontologie, et la date n'est pas absolue (« hier » est repéré pour l'événement Evenement#1). Ces améliorations seront réalisées d'ici peu afin d'obtenir des événements indépendants des textes sources.

4.3 Exploitation des annotations dans des application de veille

Les annotations insérées dans les textes peuvent ensuite être exploitées pour différents besoins. On peut par exemple extraire les événements annotés et les incertitudes associées pour alimenter automatiquement une base de connaissances de l'expert. L'affichage sur une interface peut aussi être un résultat obtenu après annotation des textes, pour une lecture enrichie des textes. Les annotations peuvent par ailleurs être exploitées par des modules de recherche d'informations sémantiques dans les textes, par exemple en recherchant tous les événements de type ASSASSINAT, ou toutes les déplacements ayant comme agent Laurent Gbagbo. Il est enfin possible de considérer l'annotation textuelle comme une première étape d'une chaîne de traitement complexe de l'information, où des modules de fusion ou d'aide à la décision peuvent se suivre afin d'aider les décideurs face à un flux d'informations important.
Dans notre contexte de veille stratégique liée à la Côte d'Ivoire, nous visons l'extraction automatique des événements repérés dans les dépêches analysées pour d'une part un affichage sur l'interface utilisateur, et d'autre part pour l'alimentation d'une base de connaissances (ontologie) qui est ensuite exploitée par un module de fusion d'informations [11]. La fusion devra permettre de capitaliser les informations concernant un même événement mais présentées dans différents textes (par exemple, la première occurrence d'un événement aura un lieu précisé, tandis qu'une autre occurrence sera associée à une date).
Notre application de veille, en cours de développement, s'appuiera sur différents services proposés par des partenaires (EADS, CEA, INRIA, …) dans le cadre du projet WebContent qui vont permettre de repérer la langue du texte, d'annoter de nouvelles entités nommées, d'afficher sur une chronologie les événements extraits, etc.
Par ailleurs, notre service d'extraction d'événements est utilisé pour l'application d'un partenaire sur la veille économique dans le domaine de l'aéronautique (EADS).

5 Validation

Nous avons mis en œuvre l'annotation automatique de textes selon des types d'événements définis précédemment, et nous avons obtenus des résultats pertinents comme cela a été montré dans l'exemple d'annotations RDF précédent. La question de l'évaluation et de la validation de notre approche est donc un sujet important, mais difficile à traiter. En effet, dans notre approche, nous partons du principe que l'utilisateur final (expert en veille stratégique) construit lui-même les patrons d'extraction associés aux différents événements recherchés, aidé par l'outil SemPlusEvent en s'appuyant sur un corpus d'apprentissage. Donc, l'évaluation de cette étape est délicate car elle dépend d'une part de la compétence et de la rigueur de l'utilisateur final pour construire un ensemble de patrons efficace, et d'autre part des événements contenus dans le corpus d'apprentissage. Par exemple, on peut imaginer un corpus d'apprentissage contenant des attentats et attaques militaires, et un nouveau corpus contenant des enlèvements et un coup d'état. Cela nous montre que l'utilisation de la phase d'apprentissage ne doit pas se limiter à un premier corpus, mais doit être toujours accessible pour permettre l'acquisition, en cours de lecture d'un nouveau corpus, de nouveaux patrons associés à de nouveaux événements. D'autre part, chaque expert s'intéresse à des informations qui n'intéressent pas forcément d'autres experts, ce qui pose le problème de la construction

d'un corpus de référence, corpus qui ne peut être créé que via une annotation manuelle, donc ayant un coût très élevé. Enfin, l'utilisation de plusieurs modules complémentaires pour le repérage des entités nommées sous toutes leurs formes, le repérage de l'incertain, le repérage des participants à chaque événement fait qu'une erreur à une étape (par exemple, erreur sur l'antécédent attribué à « Le président », repérage de la date d'annonce d'un événement au lieu de la date de l'événement, …) va entraîner l'extraction d'un événement partiellement erroné. Pour compenser ces problèmes, l'intervention de l'utilisateur final est aujourd'hui la meilleure solution, en lui donnant accès aux sources textuelles à l'origine de l'extraction des événements, et en lui permettant de valider ou modifier les caractéristiques de l'événement.
Des évaluations très restreintes de SemPlus par la STAT (Armée de Terre) [12] ont montré l'intérêt de SemPlus pour des opérationnels du renseignement. En effet, SemPlus permet de capitaliser des informations qui peuvent être réutilisées sur d'autres sujets (les patrons linguistiques associés aux événements crisogènes pour la crise en Côte d'Ivoire peuvent être utilisés pour d'autres conflits géopolitiques). Ces retours nous ont aussi permis d'améliorer notre approche, en permettant par exemple la génération de patrons plus précis. L'une des difficultés reste à évaluer le nombre de patrons qui doivent être saisis pour couvrir la majorité des occurrences d'une relation. Par exemple, lors des tests avec la STAT, 12 patrons ont été saisis pour la relation de Rencontre, 3 patrons ont été saisis pour la relation de Déplacement.
D'autres approches visant l'extraction d'événements ne proposent pas d'évaluation formelle, ou sont trop différentes pour qu'une évaluation commune soit possible. Ainsi, le système ZENON [13], qui s'appuie sur FrameNet [14] pour définir les actions (KILL, REPORT, KNOW, COMMAND, PROPOSE, EXPLODE) et entités (Company, Person, Number, Date, City, Region, River, …) à extraire d'un corpus de rapports HUMINT de la KFOR en anglais, n'a pas fait l'objet d'évaluation, faute de corpus annoté disponible et de moyens pour en créer un. Par ailleurs, l'University College Dublin [15] a aussi développé un outil pour l'extraction d'événements à partir de sources hétérogènes, mais leurs résultats sont des phrases contenant des événements, et non des événements avec leurs participants sous une forme structurée. La réalisation d'un corpus de référence pour l'extraction d'événements semble ainsi difficile à mettre en œuvre (cf [8] où en 2007 un seul concurrent a participé à l'évaluation ACE sur l'extraction d'événements), les résultats pouvant être très différents selon les besoins.

6 Conclusion

Cet article présente une application permettant d'aider des experts en veille stratégique dans leur tâche d'identification d'événements crisogènes à partir de nombreuses informations textuelles. Nous avons présenté une première classification d'événements crisogènes, qui, si elle n'est pas complète, va nous permettre d'obtenir un premier ensemble de résultats sur le cas d'étude de la Côte d'Ivoire en 2002 qui nous intéresse. Notre approche permet d'obtenir automatiquement une annotation RDF des événements grâce à l'utilisation de patrons linguistiques produits avec notre outil SemPlusEvent de capitalisation des formes textuelles associées aux événements. De plus, ces annotations sont enrichies du repérage automatique de l'incertitude exprimée dans les textes.
La mise en œuvre de notre approche s'effectue dans le cadre du projet ANR WebContent, et est sous la forme de services web qui annotent automatiquement des textes. Certaines améliorations restent à apporter à ces services, notamment un repérage fin des dates et un lien entre les lieux repérés dans les textes et les lieux définis dans l'ontologie de référence. Le résultat final que nous visons est une application de veille stratégique combinant différents services web d'annotation RDF de textes. Cette application sera composée d'une interface utilisateur permettant un affichage optimal des événements repérés dans les textes pour un expert en veille stratégique.

7 Remerciements

Le présent travail a été réalisé dans le cadre du projet ANR WebContent (www.webcontent.fr/).

8 Bibliographie

[1] **POIBEAU T**., *Extraction d'information à base de connaissances hybrides*, thèse, 2002.

[2] **GRYGLICKA E**., *Un système d'annotation des entités nommées du type personne pour la résolution de la référence*, in RECITAL 2008, 9-13 juin 2008, Avignon.

[3] **GOUJON B., FRIGIERE J.,** *Extraction of Relations between Entities from Texts by Learning Methods*, in IST-055 Specialists Meeting on "Information Fusion for Command Support", Netherlands, 2005.

[4] **GOUJON B**., *Relation Extraction in an Intelligence Context*, in LangTech 2008, 28-29 février 2008, Rome, Italie.

[5] **HEARST M. A**., 1992. *Automatic Acquisition of Hyponyms from Large Text Corpora*, in 14TH International Conference on Computational Linguistics (COLING 1992), pp. 539-545.

[6] **SILBERZTEIN M.,** INTEX : http://msh.univ-fcomte.fr/intex/

[7] **TEISSEIDRE C.,** *Développement d'un prototype pour l'extraction de relations sémantiques*, rapport interne, 2008.

[8] **ACE 2007**, http://www.itl.nist.gov/iad/894.01/tests/ace/2007/.

[9] **AUGER A., ROY J**., *Expression of Uncertainty in Linguistic Data*, in Fusion 2008, Cologne, Allemagne.

[10] **RUBIN, V., LIDDY E., KANDO N**., *Certainty Identification in Texts: Categorization Model and Manual Tagging Results*, Computing Attitude and Affect in Text: Theory and Applications, The Information Retrieval Series, Springer Netherlands, vol. 20., 2005, pp. 61-76.

[11] **LAUDY, C., GANASCIA**, J. ,2008. *Information Fusion using Conceptual Graphs: a TV Programs Case Study*. 16th International Conference on Conceptual Structures (ICCS 2008) pp. 158-165.

[12] **CYTERMANN F**., Évaluation du logiciel Sem+ dans le domaine du renseignement militaire, Rapport de stage STAT, 2005.

[13] **HECKING M**., *System ZENON – Semantic Analysis of Intelligence Reports*, in LangTech 2008, Rome, Italie.

[14] **FrameNet**, http://framenet.icsi.berkeley.edu/.

[15] **NAUGHTON M., KUSHMERICK N., CARTHY J.,** *Event Extraction from Heterogeneous News Sources*, in AAAI 2006 Workshop on Event Extraction and Synthesis, Boston.