ANNOTATION DE LA FACTUALITE D'EVENEMENTS EXPRIMEE DANS LES TEXTES

Bénédicte GOUJON
benedicte.goujon@thalesgroup.com

Thales Research & Technology France,
Campus de Polytechnique – 1, avenue Augustin Fresnel - 91767 Palaiseau Cedex (France)

Mots clefs :
annotation automatique de textes ; modèle de la factualité d'événements ; certitude ; incertitude ; discours rapportés ; point de vue ; cotation d'information
Keywords:
automatic texutal annotation ; model of event factuality ; certainty ; uncertainty ; reported discourse ; point of view ; information cotation
Palabras clave :
anotación automática de textos ; modelo de la factualidad de acontecimientos ; certeza ; incertidumbre ; discursos relatados ; punto de vista ; cotización de información

Résumé

Le travail présenté s'inscrit dans un contexte de cotation de l'information pour la veille (projet ANR Cahors). L'objectif général est de faciliter la valorisation de l'information par un veilleur en lui proposant l'extraction automatique d'événements à partir de textes. Les événements sont décrits non seulement par leurs participants (agent, lieu, date…), mais aussi par la caractérisation de leur factualité, c'est-à-dire de la réalité ou non de ces faits. Dans cet article nous détaillons notre modèle qui permet la caractérisation automatique de la factualité des événements, en utilisant des éléments de leur contexte d'énonciation : certitude ou incertitude exprimée par l'énonciateur, moment de réalisation de l'événement (passé ou à venir), négation, nom de la source locale citée, implication de la source, contact direct ou non de la source avec l'événement. Par exemple, à partir de la phrase : « Laurent Gbagbo devrait se rendre en Italie. », nous allons indiquer que l'événement associé (déplacement de Laurent Gbagbo en Italie) est exprimé avec de l'incertitude (Factuality level = Moderate).
Notre modèle de la factualité a été défini afin de permettre l'annotation automatique des textes. Nous présentons ainsi une implémentation de ce modèle, qui s'appuie sur des graphes linguistiques (automates à états finis). Grâce à ces informations relatives à leur factualité, les événements extraits automatiquement conservent les nuances exprimées dans les textes d'origine, et sont exploitables directement (pour la cotation d'informations, des besoins de veille, des chaînes de traitement d'informations textuelles...). Ces travaux se distinguent des travaux existants car ils combinent un modèle multidimensionnel (ne s'intéressant pas uniquement à l'incertain ou au discours rapporté) et une implémentation basée sur l'analyse linguistique, dans le cadre applicatif de la veille.

1 Introduction

Cet article présente un modèle et une proposition d'implémentation pour l'annotation de la factualité, c'est-à-dire de la réalité ou non d'événements décrits dans les textes. L'objectif de ce travail est de caractériser des événements extraits automatiquement à partir de textes en utilisant des caractéristiques de leur énonciation : certitude exprimée par l'énonciateur, moment de l'événement (passé ou à venir), négation, nom de la source, implication de la source. Par exemple, nous souhaitons distinguer les événements associés aux phrases suivantes : « Laurent Gbagbo s'est rendu en Italie. », « Selon notre envoyé spécial, Laurent Gbagbo doit se rendre en Italie. ». D'après la première phrase, l'événement a déjà eu lieu, tandis que selon la seconde phrase, l'événement est annoncé au futur par une source spécifiée (« notre envoyé spécial »). L'importance de ce travail a été mis en valeur par des spécialistes du renseignement de la Defence R&D Canada [1]. Ce travail s'inscrit dans le cadre du projet Cahors de cotation de l'information.

Nous présentons dans un premier temps le contexte de ce travail, qui concerne l'extraction d'événements pour la cotation de l'information. Nous décrivons ensuite les modèles existants qui sont utilisés pour des problématiques proches, ainsi que des travaux sur les discours rapportés, les modalités ou l'extraction automatique d'événements à partir de textes. Enfin, nous détaillons notre modèle d'annotation de la factualité, ainsi que son implémentation, et des utilisations possibles des annotations générées.

2 Contexte et définitions

Ce que nous appelons « événement » dans cet article est ce qui survient de façon prévue ou imprévue, qui regroupe différents participants : agent, patient, victime, lieu, destination, origine, date… et est pertinent dans un contexte de veille (stratégique, financière…). Suivant les domaines d'application, des événements sont par exemple les rencontres, les déplacements, les achats ou ventes (entre entreprises), les hausses ou baisses (cours de bourse), les accidents ou attentats…

Par factualité, nous désignons la « réalité du fait présenté ». Ce que l'on va chercher à évaluer, c'est le niveau de réalité d'un événement, en s'appuyant sur ce qui est exprimé dans le texte par son auteur (utilisation de « peut-être », conditionnel, source locale...).

2.1 La cotation de l'information

Le modèle d'annotation de la factualité décrit dans cet article est mis en œuvre dans le cadre du projet ANR Cahors (Cotation, Analyse, Hiérarchisation et Ontologies pour le Renseignement et la Sécurité). Le projet Cahors vise à organiser l'information textuelle de façon simple mais très élaborée, pour aider les services gouvernementaux à valoriser l'information en connaissance et en renseignement, dans une optique de protection du citoyen. Ce projet se focalise sur la cotation de l'information, qui correspond à l'attribution d'un ensemble d'indices de confiance. La cotation est l'une des tâches de l'étape de valorisation de l'information (entre sa capture et sa diffusion), qui est cruciale pour la veille. Aujourd'hui, au vu des masses d'informations disponibles en sources ouvertes, la cotation nécessite l'utilisation d'outils afin de soulager la charge cognitive des experts chargés de l'évaluation de l'information et de ses sources.

Dans ce projet, l'accent a été mis sur l'extraction automatique d'événements à partir de textes. Dans ce contexte, la caractérisation de la factualité, de la certitude ou incertitude exprimée dans les textes s'est avérée nécessaire, comme l'identification des sources citées dans les textes, pour une bonne évaluation de l'information par le veilleur.

2.2 Le besoin

Soit l'extrait suivant :

1. « Laurent Gbagbo s'est rendu en Italie. »

Nous disposons d'un outil d'extraction d'événements à partir de patrons linguistiques nommé SemPlusEvent (présenté dans le paragraphe 4.2.1) [4], qui va permettre d'obtenir automatiquement par l'analyse de 1. l'événement suivant : Déplacement(Agent : Laurent Gbagbo, Destination : Italie). Cet événement extrait peut ensuite être ajouté à une base de connaissances sur le domaine. La difficulté est que les événements sont rarement présentés de façon aussi factuelle. En effet, beaucoup de nuances peuvent être exprimées par un auteur, comme dans les exemples suivants :

2. « Laurent Gbagbo devrait se rendre en Italie. »
3. « Selon notre envoyé spécial, Laurent Gbagbo doit se rendre en Italie. »
4. « Finalement, Laurent Gbagbo ne s'est pas rendu en Italie. »

Le patron linguistique utilisé pour le premier exemple risque d'extraire pour ces trois exemples le même événement que pour l'exemple 1, alors que les informations exprimées sont différentes. Nous avons travaillé sur les nuances exprimées dans les textes afin de compléter les événements extraits en ajoutant des précisions : « incertitude » et « événement à venir » pour l'exemple 2, « information rapportée » et « événement à venir» pour l'exemple 3, « négation » pour l'exemple 4. L'objectif final est de rendre l'événement extrait de son contexte (par exemple ajouté dans la base de connaissance) le plus proche de ce qui est exprimée dans le texte source. Nous avons besoin d'un modèle regroupant ces différentes informations sur le contexte d'énonciation d'un événement, ainsi qu'une méthode d'implémentation de ce modèle, afin de mettre en œuvre l'annotation automatique relative à ce modèle.

3 État de l'art

Notre travail se trouve à la croisé de travaux de linguistique sur les modalités, la négation et les discours rapportés. L'utilité de l'identification de l'incertain exprimé dans les textes pour la veille, en utilisant des connaissances linguistiques, a été mis en avant par des spécialistes du renseignement canadien [1], mais ils ne décrivaient ni nouveau modèle ni implémentation. Nous présentons ici deux modèles existants qui visent des problématiques semblables aux nôtres, ainsi que différents travaux de linguistique autour du discours rapporté, des modalités et de l'évidentialité.

3.1 Modèle de Rubin

Rubin et al [10] ont proposé un modèle de catégorisation de la certitude, basé sur quatre dimensions : « level », « perspective », « focus » et « time ». Le schéma ci-dessous précise ces dimensions.

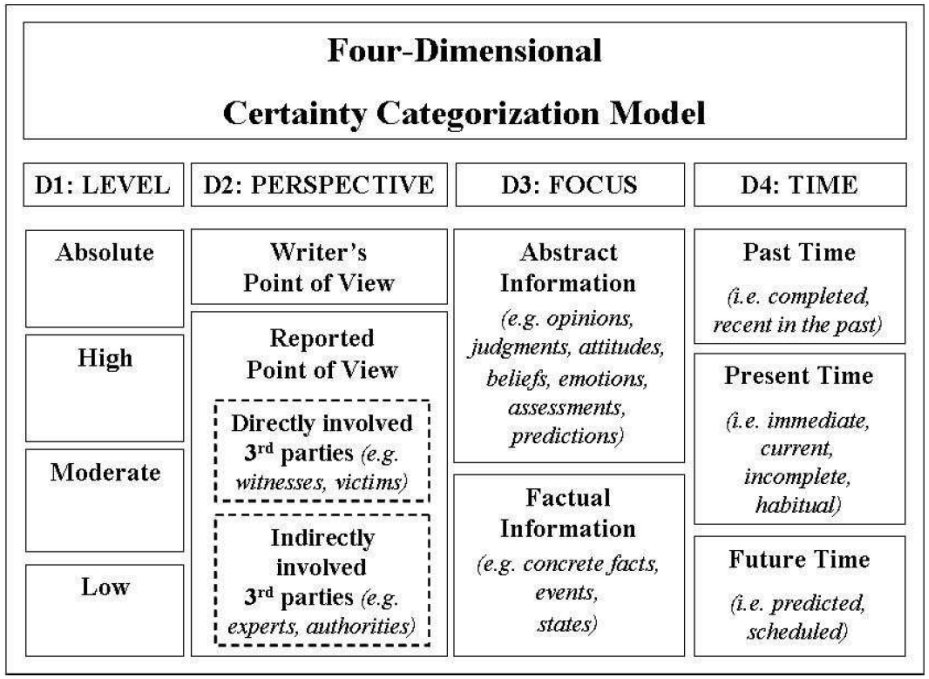

Figure 1 : Modèle de Rubin.

Notre modèle s'est inspiré de celui-ci, en l'adaptant à notre problématique. Ainsi, le modèle de Rubin détecte le niveau de certitude, qui est exprimé par un énonciateur, tandis que dans notre approche centrée sur les événements, nous détectons le niveau de factualité. Ces deux niveaux sont liés mais décrivent des caractéristiques différentes. Par ailleurs, le modèle de Rubin ne tient pas compte de la négation, qui est quelque chose d'important à repérer dans notre approche afin d'identifier si un événement n'a pas eu lieu. De plus, ce modèle propose une dimension focus permettant de distinguer les informations factuelles ou abstraites. Dans notre approche, nous nous intéressons exclusivement aux événements, qui sont des informations factuelles. Enfin, ce modèle n'a pas fait l'objet de mise en œuvre dans un module d'annotation automatique, il est uniquement utilisé pour annoter manuellement des corpus.

3.2 Modèle de Saurí

Dans Saurí [11], les auteurs proposent un modèle pour l'annotation de la factualité des événements d'un corpus. Ce modèle est utilisé pour l'annotation manuelle du corpus FactBank (208 documents, plus de 8 000 événements), en complément d'annotations produites à partir de TimeML. Le tableau suivant illustre les catégories qui sont annotés pour caractériser la factualité des événements.

	Positive	Negative	Underspecified
Certain	Fact: <CT,+>	Counterfact: <CT,−>	Certain but unknown output: <CT, u>
Probable	Probable: <PR,+>	Not probable: <PR,−>	(NA)
Possible	Possible: <PS,+>	Not certain: <PS,−>	(NA)
Underspecif.	(NA)	(NA)	Unknown or uncommitted: <U,u>

Figure 2 : Valeurs de factualité selon Saurí.

Pour mieux comprendre ce modèle, le certain sous-spécifié noté <CT,u> est illustré par « *John knows whether Mary came.* ». Dans le modèle TimeML, la majorité des verbes et des noms prédicatifs expriment des événements, comme le montre la phrase suivante, contenant 8 événements, issus du corpus TimeBank :

13 The move seemed$_{e83}$ aimed$_{e84}$ at heading$_{e85}$ off more trouble$_{e86}$ with Iran , which had condemned$_{e87}$ Iraq 's invasion$_{e88}$ of Kuwait on Aug. 2 but also criticized$_{e90}$ the multinational force dispatched$_{e91}$ to Saudi Arabia .

Dans notre approche, la notion d'événement est plus restreinte, se limitant aux événements qui intéressent l'utilisateur. De plus, la distinction probable – possible, qui est au cœur de ce modèle, ne nous semble pas utile dans notre approche, où l'on va plutôt chercher à savoir si un événement a eu lieu ou non, ou avec quel degré de certitude l'auteur le présente.

3.3 Travaux sur les discours rapportés, les modalités et l'évidentialité

Le discours rapporté, l'un des moyens d'exprimer un détachement vis-à-vis du propos, a été souvent étudié, mais pour des objectifs différents du nôtre. Par exemple, chez Battistelli et Chagnoux [2], le discours rapporté est exploité pour représenter la dynamique énonciative et modale de textes, c'est-à-dire pour repérer les relations discursives entre différentes propositions de textes. Différents référentiels sont associés aux propositions : référentiel énonciatif, référentiel possible, référentiel mental. L'objectif est d'expliciter la structure hiérarchique d'un texte, mais pas d'attribuer des valeurs de factualité ou de certitude combinée à la source comme nous souhaitons le faire.

Les travaux sur les modalités, tels que ceux de B. Pottier[1], distinguent différentes catégories de modalités, qui sont étudiées selon leurs interactions, leurs points communs ou leurs spécificités. Dans notre approche, nous avons principalement exploiter la catégorie modale épistémique, liée au savoir et à la certitude, qui nous semble la plus pertinente par rapport au problème que nous devons traiter.

L'évidentialité, qui est présentée en détail par Dendale et Coltier [3], concerne le mode d'accès à une information. Dans cet article, trois cas d'évidentialité sont présentés. L'évendialité par la perception directe, qui correspond aux situations où l'énonciateur a été en contact avec ce qu'il décrit

[1] Synthèse présentée dans Ouattara A., Modalités et verbes modaux dans les écrits de Bernard Pottier, Les Verbes Modaux, Cahiers Chronos 8 (2001) p.1-16.

(il a vu ce qui s'est passé). L'évidentialité par la reprise, qui correspond aux discours rapportés. Enfin, l'évidentialité par l'inférence, à partir d'indices, qui est par exemple introduite par le verbe « penser ». Cette notion est en rapport direct avec notre problématique, et se retrouve à travers des valeurs de plusieurs caractéristiques de notre modèle (chaque cas d'évidentialité est mis en valeur dans la description de notre modèle ci-après).

4 Description de notre approche

Le modèle que nous présentons ici a pour objectif d'être adapté pour l'annotation automatique de la factualité dans les textes, en complément de l'annotation automatique d'événements dans un contexte de veille. Cette contrainte de combinaison du modèle et de son implémentation distingue notre approche des travaux de modélisation de Rubin et de Saurí, qui ont proposé leurs modèles indépendamment de leurs mise en œuvre éventuelle.

4.1 Le modèle pour l'annotation de la factualité exprimée dans les textes

Le modèle que nous avons défini s'appuie sur trois dimensions (obligatoires) et trois marqueurs (optionnels).

Factuality level – D	Absolute	High	Moderate	Low
Time – D	Past	Present	Future	
Source – D	Writer	Source Name		
Negation – M				
Primary Source Perspective – M				
Commitment – M	Low	High		

Figure 3 : Le modèle d'annotation de la factualité (D : dimension, M : marqueur).

Les dimensions et marqueurs de ce modèle sont présentés en détail ci-après. Nous avons défini ce modèle afin de regrouper l'ensemble des caractéristiques du contexte d'énonciation d'un événement. Le but est de permettre à un utilisateur final, découvrant l'événement sous forme structurée et isolé du contexte textuel, de déduire si l'événement en question a déjà eu lieu, a peut-être eu lieu ou aura lieu. Pour cette déduction, l'utilisateur final devra utiliser ses propres connaissances sur la fiabilité des sources (que nous n'envisageons pas d'identifier automatiquement). Notre modèle est plus

complet et plus adapté pour répondre à notre problématique d'identification de la factualité d'un événement à l'aide de différentes caractéristiques que les autres modèles présentés précédemment.

4.1.1 Level

La dimension « *Level* » correspond dans notre modèle au niveau de factualité exprimé. Ici, le terme « factualité » est pris dans le sens « réalité du fait présenté ». Ce que l'on cherche à évaluer, c'est le niveau de réalité d'un événement, en s'appuyant sur ce qui est exprimé dans le texte par son auteur. Ainsi, si aucune nuance n'est exprimée (exemple 1 initial), « *Level* » a la valeur « *Absolute* » qui est sa valeur par défaut (on suppose que l'auteur est convaincu de la réalité de l'événement qu'il rapporte). Si une forte certitude est exprimée (sûrement, sans aucun doute, affirmer…) quant à la réalité de l'information, « *Level* » a la valeur « *High* ». Si une faible certitude est exprimée (peu probable, incertain…), « *Level* » a la valeur « *Low* ». Dans les autres cas, où une incertitude moyenne est exprimée (via « peut-être », l'utilisation du conditionnel, « il paraît que »…), « *Level* » est associé à la valeur « *Moderate* ».
Ce nombre de niveaux, qui s'inspire du modèle de Rubin, permet une compréhension de chaque valeur. En revanche, il ne permet pas de distinguer toutes les nuances qui peuvent être exprimées (probable et fort probable sont associés à « *High* »). Contrairement au modèle de Rubin, cette dimension exprime la factualité, et non la certitude, ainsi « Il est inimaginable que Laurent Gbagbo se rende en Italie. » sera associé à « *Low* », car cette phrase affirme le peu de chance que l'événement ait lieu.

4.1.2 Time

La dimension « *Time* » ne correspond pas uniquement au temps du verbe qui est au cœur du patron d'identification d'un événement. Elle correspond plus précisément au moment auquel se passe l'événement par rapport au temps de l'énonciation. Ainsi, « se rendra », « doit se rendre » « se rend demain » ou « se rendrait » correspondent à la valeur « *Future* ». Pour repérer la valeur « *Present* », nous utilisons des expressions telles que « en ce moment », « actuellement » ou « être en train de ». La valeur par défaut est « *Past* ».

4.1.3 Source(s)

La dimension « *Source(s)* » permet de marquer un discours rapporté via l'identification de la source d'une information. Par défaut, sa valeur est « *Writer* », car c'est l'auteur du texte qui est la première source. Si une source locale est précisée : « Selon notre envoyé spécial… », « Le porte-parole a annoncé… », son contenu est récupéré dans la dimension « *Source(s)* ». La source n'est pas forcément une personne, elle peut aussi être un média, un document, une organisation politique... Lorsque la valeur de cette dimension est différente de « *Writer* », on est dans le cas d'un discours rapporté, ce qui correspond à l'évidentialité par la reprise chez Dendale et Coltier [3].
Quand on se trouve dans un cas de sources imbriquées, comme dans : « … a déclaré une porte-parole de la compagnie, cité par l'agence Reuters. », cette dimension contient toutes les sources qui sont spécifiées dans la phrase, de façon ordonnée (de la moins impliquée à la plus impliquée) : « agence Reuters » > « une porte-parole de la compagnie ».

4.1.4 Negation

Le marqueur de négation permet d'identifier quand un événement est présenté accompagné d'une tournure négative (cf exemple 4). La négation ne fait pas toujours l'objet d'analyse spécifique dans les approches d'extraction automatique d'informations à partir de textes, et dans notre cas elle est nécessaire à une bonne compréhension de l'événement extrait de son contexte. On aurait pu choisir de ne pas tenir compte d'un événement associé à une négation, mais il nous semble plus pertinent de le prendre en compte enrichi d'un marqueur « *Negation* », afin par exemple de permettre à un utilisateur final de suivre jusqu'au bout un événement annoncé qui est finalement annulé ou démenti ou de comparer des propos contradictoires de différentes sources.

4.1.5 Primary Source Perspective

Le marqueur « *Primary Source Perspective* » permet de spécifier si la source a été en contact direct avec l'événement décrit, quand cela est explicité. Si l'on a « Notre envoyé spécial a vu Laurent Gbagbo à Rome », le marqueur « *Primary Source Perspective* » sera activé. Dans les autres cas, par exemple « Notre envoyé spécial pense que Laurent Gbagbo était à Rome », ce marqueur ne sera pas utilisé. Ce marqueur est en relation avec l'évidentialité par la perception directe de Dendale et Coltier [3].

4.1.6 Commitment

Le degré d'implication de la source dans ce qui est décrit est exprimé dans le marqueur « *Commitment* ». Lorsque la source est peu impliqué : « Laurent Gbagbo pense que… », sa valeur est « *Low* ». Ce cas correspond à l'évidentialité à partir d'indices, telle que décrite par Dendale et Coltier. Lorsque la source est très impliquée : « Laurent Gbagbo affirme que… », sa valeur est « *High* ». Quand plusieurs sources sont identifiées, ce marqueur s'applique pour la source la plus impliquée uniquement.

4.1.7 Modèles complémentaires

Dans notre modèle d'annotation de la factualité, nous ne cherchons pas à caractériser les opinions, sentiments, jugements ou émotions de la source par rapport à ce qu'elle rapporte. Ces éléments ne sont généralement pas exprimés par l'auteur de l'article de presse, mais sont très présents dans les cas de discours indirects, comme dans « "La scène est aussi effroyable que vous pouvez l'imaginer" , a commenté le maire de la ville » ou « Même le quotidien Le National… a condamné cette descente ». Nous avons choisi de centrer notre modèle sur la réalité ou non d'un événement qui est décrit dans un texte, nous nous sommes éloigné de la première version de ce modèle qui cherchait à caractériser la certitude ou incertitude exprimée par la source. Les verbes exprimant des opinions ou jugements tels que « condamner » ou « juger » sont utilisés comme marqueurs de discours indirects exprimés par une source locale. En ce qui concerne l'identification des opinions ou sentiments, certains travaux ont déjà produits des modèles dédiés (cf Mathieu [9]), un projet nommé DOXA[2] se focalise actuellement sur ce sujet, et leurs résultats pourraient être utilisés en complément de notre modèle.

4.2 Le module d'annotation automatique

4.2.1 Annotation des événements avec l'outil SemPlusEvent

Le repérage d'événements dans les textes est réalisé grâce à notre outil SemPlusEvent [4], dont l'objectif est d'une part de faciliter la capture de patrons syntaxico-sémantiques associés à des types d'événements, et d'autre part d'appliquer ces patrons afin d'extraire des instances d'événements à partir de textes. Pour construire les patrons linguistiques, la première étape consiste à repérer les participants de ces événements. Ces participants sont principalement désignés par des entités nommées : lieux, personnes, organisations, dates, … L'étape suivante consiste à lister les formes textuelles que peut prendre chaque événement. 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 … », où X et Y sont des entités nommées de type Personne. Pour chacune des expressions relatives à un type d'événement, l'utilisateur définit via l'interface de l'outil SemPlusEvent le rôle de chaque entité (patient, agent, lieu…). Enfin, la dernière étape consiste à introduire des généralisations, en remplaçant les instances d'entités nommées (ex : Laurent Gbagbo) par leur catégorie (ici Personne), et les expressions verbales telles que « a été tué » par l'ensemble des formes passives de tuer : « vient d'être tué », « ont été tué », « aurait été tué », etc. Au final, des patrons syntaxico-sémantiques sont générés automatiquement par l'outil SemPlusEvent (figure 4), sans que l'utilisateur n'ait eu besoin de connaissances linguistiques spécifiques.

[2] Projet DOXA du pôle Cap Digital : https://www.projet-doxa.fr/

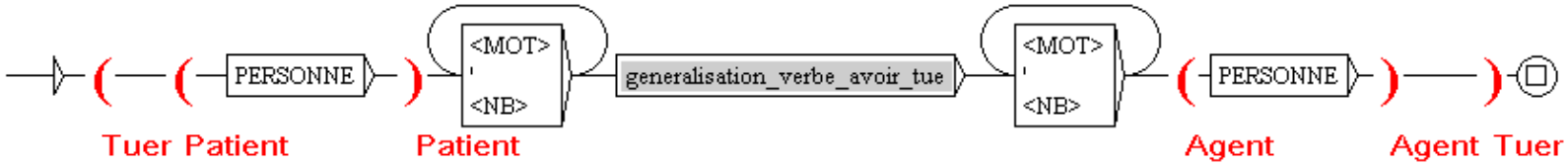

Figure 4 : Exemple de patron lexico-syntaxique généré par SemPlusEvent.

Ces patrons, premiers résultats de SemPlusEvent, sont ensuite appliqués sur des textes pour permettent l'identification de nouvelles instances d'événements, qui mettent en relation différentes entités nommées. La copie d'écran suivante (figure 5) montre l'affichage de deux instances d'événements de type Rencontre entre deux personnes, qui ont été repérées dans des textes. Des couleurs permettent de distinguer les types des entités nommées désignées dans les phrases.

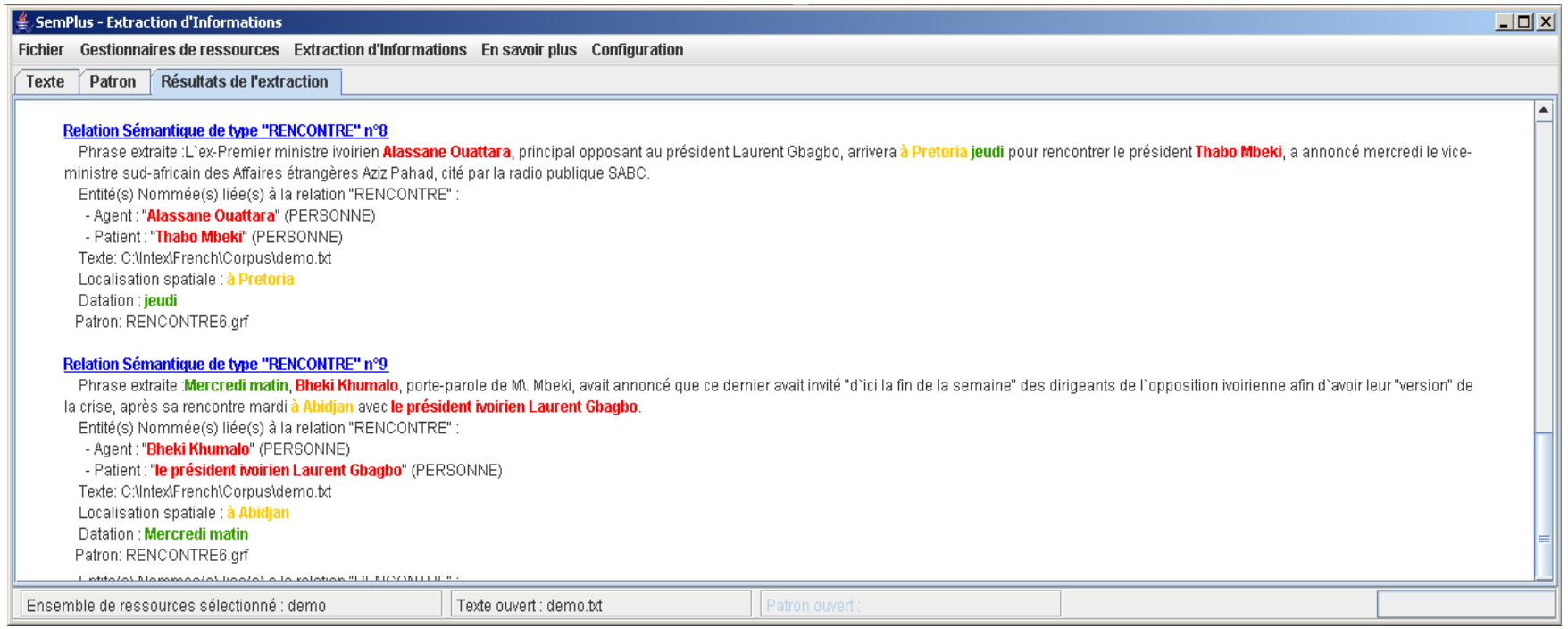

Figure 5 : Affichage des événements extraits avec SemPlusEvent.

Sur cet interface, le premier résultat concerne une relation de type RENCONTRE entre « Alassane Ouattara » (type : Personne, rôle : Agent) et « Thabo Mbeki » (type : Personne, rôle : Patient), issue du texte demo.txt, et associée au lieu « à Pretoria » et à la date « jeudi ». Ce résultat a été obtenu grâce au patron RENCONTRE6.grf appliqué à la phrase « L'ex-Premier ministre ivoirien Alassane Ouattara, principal opposant au président Laurent Gbagbo, arrivera à Pretoria jeudi pour rencontrer le président Thabo Mbeki, a annoncé mercredi … ». Toutes ces informations permettent à un utilisateur final de visualiser clairement les relations extraites, de pouvoir facilement retrouver les sources textuelles de ces relations et éventuellement de corriger des patrons qui auraient produit des résultats erronés.

4.2.2 Détails de l'implémentation en cours

Le module d'annotation de la factualité exprimée dans les textes en français s'appuie, dans sa version actuelle, sur 34 graphes écrits avec Unitex[3], impliquant 96 mots (noms, verbes, adjectifs…) ou expressions, organisés selon les dimensions et marqueurs spécifiés ci-dessus. Ces graphes sont actuellement appelés via le serveur sémantique d'Arisem. L'intégration de ce module d'annotation à notre outil d'extraction d'événements SemPlusEvent est en cours, et sera suivie d'une évaluation.
Le graphe présenté ci-après (Figure 6) annote la source et la phrase complète quand celle-ci contient l'un des verbes suivants qui expriment un fort engagement de la source : nier, contester, affirmer, confirmer, certifier, jurer, assurer. L'expression verbale est généralisée, afin de repérer toutes les formes conjuguées possibles des verbes : « vient d'affirmer », « devrait confirmer », « a fortement nier »…

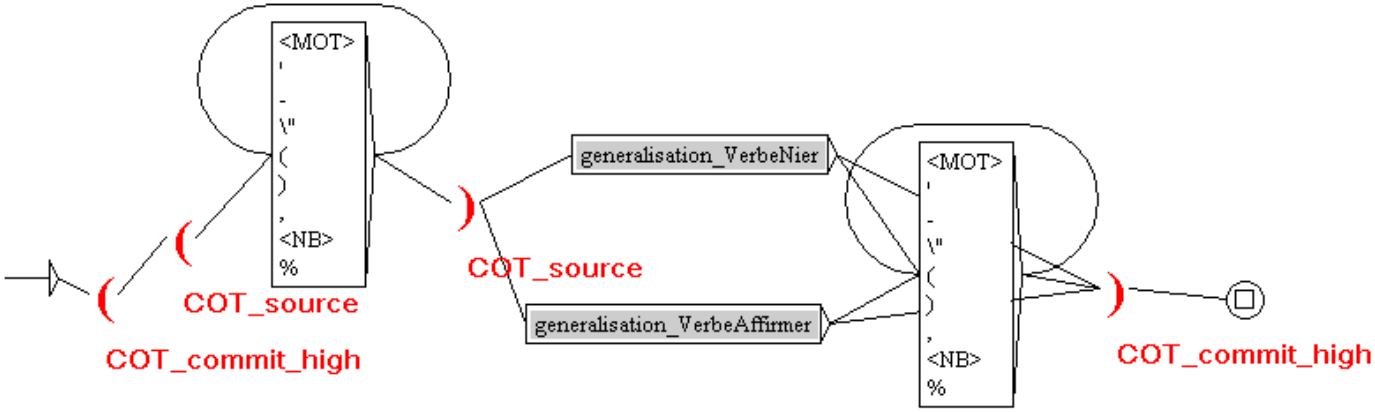

Figure 6 : Exemple de graphe repérant le fort engagement d'une source (COT = cotation).

Concernant la dimension « *Level* », quand plusieurs niveaux de factualités sont exprimés conjointement, par exemple dans : « Laurent Gbagbo devrait vraisemblablement se rendre en Italie. », où le conditionnel est associé à « *Moderate* », tandis que vraisemblablement est associé à « *High* », on fait le choix de conserver la valeur extrême (« *High* » ou « *Low* »). La négation combiné à un marqueur de la valeur « *High* » produit une valeur « *Low* » (et inversement), comme dans « Laurent Gbagbo ne se rendra probablement pas en Italie. ».
Pour la dimension « *Time* », des marqueurs de présent ou de futur sont recherchés (temps des verbes, « demain », « actuellement »…), la valeur « *Past* » étant attribuée par défaut..

[3] IGM : http://www-igm.univ-mlv.fr/~unitex/

La dimension « *Source(s)* » est traitée à un niveau inter-phrastique, car certains discours rapportés (notamment ceux encadrés par des guillemets) peuvent s'étendre sur plusieurs phrases. La source est repérée comme étant suivie d'un verbe introduisant un discours rapporté (dire, révéler, revendiquer…), ou elle est précédée par « selon ». On envisage l'ajout d'une valeur « *Unidentified* » si le module chargé de l'annotation automatique ne peut retrouver la source citée dans le texte, afin de ne pas perdre l'aspect discours indirect. On envisage d'utiliser des connaissances identifiant différents niveaux de sources (sources non officielles : proches de témoins ou victimes ; sources officielles : policiers ou porte-parole ; médias : agences de presse, journalistes, journaux, chaînes de télévision…), afin d'ordonner les sources.

4.2.3 Évaluation d'une première implémentation

L'implémentation d'une première version de notre modèle, initialement orienté annotation de l'incertitude et proche de celui de Rubin, a été présentée dans [4]. Cette implémentation avait pour objectif de montrer la faisabilité de l'annotation automatique de ce genre de modèle à l'aide de graphes. Ce premier modèle a été implémenté dans un service web produisant des annotations au format RDF des événements et de l'incertitude exprimée dans les textes, dans le cadre du projet ANR WebContent[4]. L'annotation de l'incertitude étant appliquée avant l'annotation des événements, cela nous a permis de réaliser une première évaluation centrée sur ce modèle de l'incertain, indépendamment des événements extraits. Ainsi, parmi les limites de cette première implémentation, on associait initialement le niveau de certitude Moderate à chaque discours rapporté, puisque ces derniers expriment une non prise en charge par l'auteur du texte des propos rapportés. Or, les discours rapportés peuvent aussi être utilisés pour de simples citations, comme dans : « "Nos pensées et nos prières vont aux familles et aux amis touchés par cette tragédie", a ajouté le président… » ou pour citer des mots trop porteurs de point de vue, comme dans « … il n'y a eu "aucune irrégularité" dans son élection. ». Il n'est donc pas pertinent d'associer systématiquement de l'incertitude à un discours rapporté.
L'implémentation du modèle qui est présenté ici n'a pas encore été évalué. Son évaluation nécessite la construction (très coûteuse) d'un corpus assez volumineux de référence, qui serait annoté manuellement selon les différents éléments du modèle, et sa comparaison avec le même corpus analysé automatiquement.

4.3 Utilisation envisagée des annotations

Les annotations produites peuvent être utilisées dans différents contextes.
Elles peuvent par exemple être utilisées par des personnes, dans le cadre d'une veille, qui souhaitent connaître de façon précise une situation en cours. Pour cela, ces personnes vont utiliser nos outils d'extraction d'événements afin de traiter un maximum de textes, et elles pourront vérifier la réalité d'événements extraits automatiquement en prenant connaissance des nuances exprimées dans leurs contextes d'énonciation via la factualité associée à chaque événement. Pour confirmer la réalité d'un événement, dans un contexte de cotation de l'information (projet Cahors), l'utilisateur aura besoin de ses connaissances personnelles sur la fiabilité des sources identifiées.
Les annotations de factualité des événements peuvent aussi être utilisées dans le cadre de chaînes de traitement de l'information textuelle, par des outils de fusion d'informations par exemple. Ces outils vont s'appuyer sur les valeurs des différents marqueurs et dimensions associés aux différentes occurrences d'un même événement, afin de savoir quelle occurrence contient une valeur « prioritaire », plus proche du fait réel, par rapport à une autre. Un événement présenté au futur et un événement présentant les mêmes caractéristiques mais publié ultérieurement et présenté au passé, pourront être fusionnés en un seul événement.

[4] www.webcontent.fr/

4.4 Les autres approches d'extraction d'informations

Il existe différents outils permettant d'extraire des événements à partir de textes, tels que Zenon, décrit par Hecking [6], qui s'appuie sur GATE pour l'extraction d'événements (KILL, REPORT, KNOW, COMMAND, PROPOSE, EXPLODE) à partir de rapports de la KFOR, mais ils ne s'intéressent pas à l'extraction de la factualité exprimée dans le contexte des événements qui sont extraits.
Dans les campagnes ACE (Automatic Content Extraction)[5] du NIST, l'identification d'attributs tels que les modalités ou la polarité d'événements était présente en 2007, mais a été supprimée en 2008, un seul candidat (BBN Technologies) ayant tenté cette tâche en 2007. Ce genre de tâche n'apparaît pas dans la campagne plus récente TAC. L'identification de la factualité peut avoir un intérêt limité dans les cas où l'on veut uniquement extraire des événements à partir de textes, sans chercher à exploiter réellement les résultats obtenus ou dans les cas où l'on s'intéresse spécifiquement aux discours rapportés,. Dans notre cas, où les événements extraits sont amenés à être utilisés, la caractérisation de la factualité nous semble nécessaire, comme nous l'avons montré en début d'article.
D'autres outils traitent un seul des aspects de notre modèle. Par exemple, Krestel et al. [8] visent l'annotation automatique des discours rapportés et de leurs sources, pour l'analyse d'articles de presse, en anglais. Pour cela, leur module utilise une liste de verbes d'introduction de discours rapportés (believe, accuse, note, add…) et sur des patrons s'appuyant sur ces verbes. Ce module, développé à partir de GATE, ne s'intéresse pas à la factualité d'événements pouvant être décrits dans des discours rapportés.

5 Conclusion

Nous avons présenté dans cet article notre modèle de caractérisation de la factualité des événements qui sont décrits dans les textes. L'objectif est ici de savoir si un événement a eu lieu ou s'il devrait avoir lieu, en exploitant notamment les sources d'informations citées par l'auteur principal (via des discours indirects), leur implication dans ce qu'elles rapportent, leurs certitudes ou incertitudes.
Ce travail se situe dans le contexte de la cotation de l'information pour la veille. Notre modèle, qui contient trois dimensions et trois marqueurs optionnels, a été présenté en détails et mis en rapport avec des modèles proches existants. L'intérêt de notre travail est qu'il inclut une implémentation du modèle, afin de produire des annotations automatiques. Cette mise en œuvre de ce modèle devra faire l'objet d'une évaluation. Le modèle pourra évoluer s'il apparaît que certaines valeurs ne peuvent être distinguées automatiquement. La prise en compte de la subjectivité des sources, vis-à-vis des informations décrites (opinions, jugements de valeurs, sentiments) ou leur intervention, action dans les informations présentées, pourrait se faire dans un autre modèle complémentaire, qui se focaliserait non plus sur les événements mais sur les sources.

6 Remerciements

Je remercie les partenaires du projet Cahors qui ont participé à des réunions sur le sujet abordé ici et qui ont permis, grâce à leurs remarques pertinentes, de faire évoluer le modèle, en particulier Nicolas Dessaigne et Aurélie Migeotte d'Arisem, Thomas Delavallade et Philippe Capet de Thales Communications, Gloria Origgi et Hadi Ba de l'Institut Jean Nicod.

7 Bibliographie

[1] AUGER A., ROY J. (2008). Expression of Uncertainty in Linguistic Data, Actes de Fusion 2008, Cologne, Allemagne.

[2] BATTISTELLI D., CHAGNOUX M. (2007). Représenter la dynamique énonciative et modale de textes, Actes de TALN 2007, Toulouse, p.23-32.

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

[3] DENDALE P., COLTIER D. (2003). Point de vue et évidentialité. *Cahiers de praxématique*, 41, p.105-129.

[4] GOUJON B. (2009). Uncertainty Detection for Information Extraction, Actes de RANLP 2009, Borovets, Bulgaria.

[5] GOUJON B. (2009). Annotation d'événements dans les textes pour la veille stratégique, in VSST 2009, Nancy.

[6] HECKING M. (2008). System ZENON – Semantic Analysis of Intelligence Reports, in LangTech 2008, Roma, Italy.

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

[8] KRESTEL R., BERGLER S., WITTE R. (2008). Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles, in LREC 2008, Marrakech, Maroc.

[9] MATHIEU Y. Y. (2008). Navigation dans un texte à la recherche des sentiments. *Linguisticae Investigationes*. 31 :2, pp. 313-322.

[10] RUBIN V. L., LIDDY E. D., KANDO N. (2005). 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. pp. 61-76.

[11] SAURÍ R., PUSTEJOVSKY J. (2008). From structure to interpretation: A double-layered annotation for event factuality, Actes de LREC 2008, Marrakech, Maroc.

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