Représentation évènementielle des déplacements dans des dépêches épidémiologiques

Manal El Zant[1], Jean Royauté[1], Michel Roux[1]
(1) LIF / Université de la Méditerranée, 27 Bd Jean Moulin, 13005 Marseille
el.zant@medecine.univ-mrs.fr, jean.royaute@lif.univ-mrs.fr,
michel.roux@medecine.univ-mrs.fr

Résumé. La représentation évènementielle des déplacements de personnes dans des dépêches épidémiologiques est d'une grande importance pour une compréhension détaillée du sens de ces dépêches. La dissémination des composants d'une telle représentation dans les dépêches rend difficile l'accès à leurs contenus. Ce papier décrit un système d'extraction d'information utilisant des cascades de transducteurs à nombre d'états fini qui ont permis la réalisation de trois tâches : la reconnaissance des entités nommées, l'annotation et la représentation des composants ainsi que la représentation des structures évènementielles. Nous avons obtenu une moyenne de rappel de $80,93\%$ pour la reconnaissance des entités nommées et de $97,88\%$ pour la représentation des composants. Ensuite, nous avons effectué un travail de normalisation de cette représentation par la résolution de certaines anaphores pronominales. Nous avons obtenu une valeur moyenne de précision de $81,72\%$ pour cette résolution.

Abstract. The representation of motion events is important for an automatic comprehension of disease outbreak reports. The dispersion of components in this type of reports makes it difficult to have such a representation. This paper describes an automatic extraction of event structures representation of these texts. We built an information extraction system by using cascaded finite state transducers which allowed the realization of three tasks : the named entity recognition, the component annotation and representation and the event structure representation. We obtained a recall of $80,93\%$ for the named entity recognition task and a recall of $97,88\%$ for argument representation task. Thereafter, we worked in anaphoric pronouns resolution where we obtained a precision of 81.83%.

Mots-clés : Sous-langage, représentation évènementielle, extraction d'information, structure prédicative, structure predicate-arguments.

Keywords: Sublanguage, event structure representation, information extraction, predicative structure, predicate-arguments structure.

1 Introduction

Dans le cadre de la veille sanitaire, le projet Epidém*IA* a pour but de bâtir un système d'aide à la décision en utilisant les caractéristiques des épidémies décrites dans les dépêches épidémiologiques. Ce projet comporte deux principaux modules : (i) un module de traitement automatique de la langue naturelle (notre contribution dans ce papier) pour obtenir une représentation évènementielle des textes. Cette représentation doit tenir compte de l'agrégation des évènements et

de leur localisation spatio-temporelle ; (ii) un outil exploitant un langage formel de représentation des connaissances (*STEEL*) adapté à la problématique des informations épidémiologiques et tenant compte de la propagation des évènements dans le récit (Chaudet, 2004).
Notre objectif a été le développement d'un système d'extraction d'information permettant la représentation évènementielle du contenu de ces dépêches. Cette représentation doit tenir compte des différentes formes d'informations qui composent les évènements ainsi que de leur localisation spatiale et temporelle. Pour que nos sorties puissent être utilisées par la suite par le module *STEEL*, nous adoptons pour ses sorties une représentation prédicative logique dans laquelle la quantification existentielle est implicite.

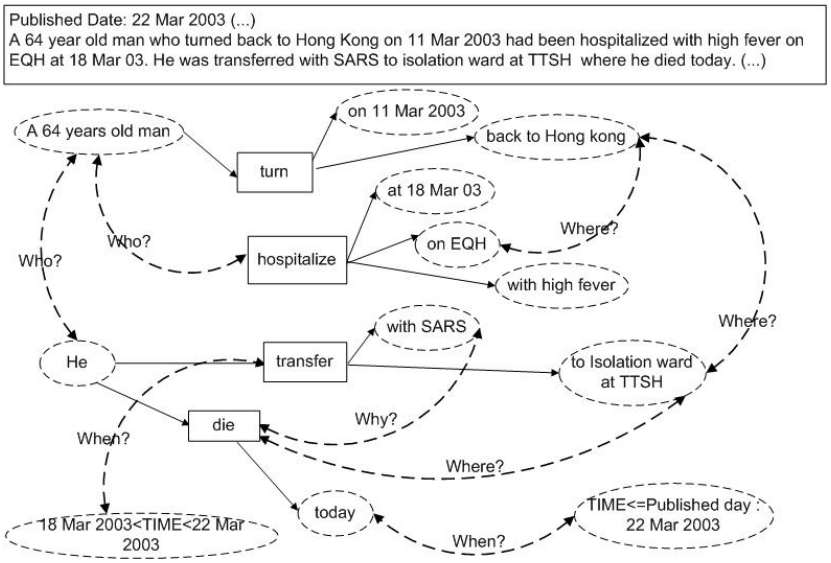

FIG. 1 – Représentation évènementielle des données épidémiologiques.

La complexité du contenu des dépêches rend difficile une telle représentation. Cette complexité est liée aux relations d'inclusion entre les différents évènements et à la dispersion des composants d'un évènement dans plusieurs autres évènements (Huttunen *et al.*, 2002a; Huttunen *et al.*, 2002b). La figure 1 montre les informations détaillées des différents évènements d'un exemple de dépêches. Le contenu des nœuds représentent les formes de surface des composants de cette dépêche (*personne concernée*, *cause*, *lieu* et *temps*). Les flèches en traits pleins représentent les relations entre verbes et composants, les flèches en pointillés permettent de récupérer les composants omis. La tâche de normalisation est conçue pour trouver les composants omis ou incomplet (autrement dit, répondre aux questions de la figure 1 : *when ?*, *where ?*, *who ?* et *why ?*). Notons ici que les deux questions *when ?* et *where ?* ont deux niveaux de traitement : (i) le cas où une localisation est totalement absente et pour laquelle la question porte sur une relation directe avec le verbe (ici, la question *when ?* pour le verbe *transfer* et la question *where ?* pour le verbe *die*) ; (ii) le cas où une localisation existe mais est incomplète : la question nécessite alors un enrichissement d'information sur cette localisation (ici, la question *when ?* pour l'expression *today* pour laquelle nous précisons la date et la question *where ?* pour les deux localisation *EQH* et *isolation ward at TTH*).
Dans la section 2, nous présentons la structure globale de notre système. Puis la section 3 détaille les différentes étapes de l'analyse locale qui consiste en une reconnaissance, annotation et représentation prédicative logique des composants du sous langage des dépêches. Nous utilisons à cette étape une typologie des verbes afin d'avoir une représentation des structures évènementielles. Finalement, la section 4 est consacrée à résoudre certaines anaphores pronominales, ce qui permet, grâce à un ensemble de règles de normalisation, de compléter des informations incomplètes.

2 Description du système

Une architecture générale des systèmes d'extraction d'information (SEI) est proposée par (Grishman, 1997). Un SEI traitant la surveillance des épidémies a en plus pour objectif de décrire l'évolution de la population infectée. Les systèmes de ce domaine font partie des systèmes de Bio-sécurité (*IFE-Bio*[1]) (Hirschman *et al.*, 2001; Sears & Cross, 2001). Dans cette perspective, deux systèmes permettent d'analyser les dépêches épidémiologiques : le système *Proteus-BIO* (Grishman *et al.*, 2002a; Grishman *et al.*, 2002b) et le système *MiTAP*[2] (Damianos *et al.*, 2002; Damianos *et al.*, 2003; Damianos *et al.*, 2004). L'objectif de l'extraction est de pouvoir identifier pour chaque dépêche le lieu et le temps de l'épidémie, le nom de l'épidémie, le nombre et le type de personnes concernées en précisant s'il s'agit de personnes mortes ou malades. Les projets *Proteus-Bio* et *MiTAP* utilisent respectivement le système d'extraction d'information *NYU* (Grishman, 1995) et le système *Alembic* (Aberdeen *et al.*, 1995). Avec une adaptation aux exigences du domaine épidémiologique, ces deux systèmes ont apporté des réponses aux problèmes d'extraction des scénarios d'évènements de surveillance en respectant les attentes des projets *IFE-Bio*.

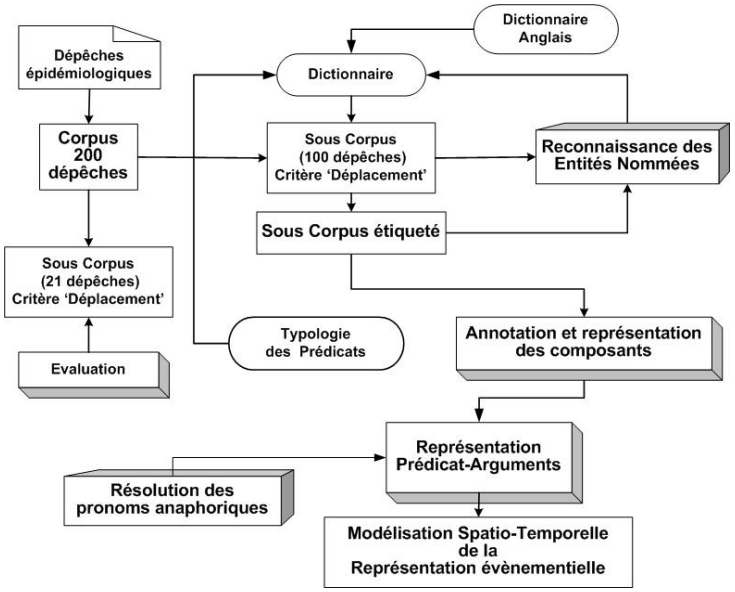

FIG. 2 – Architecture de notre Module

Notre approche est similaire aux projet *Proteus* et *MiTAP* par l'utilisation de ces dépêches pour analyser la situation épidémiologique. Par contre, notre but n'est pas la surveillance de l'évolution du nombre de cas d'une épidémie. Nous avons orienté notre recherche sur le phénomène du déplacement humain qui est très important en épidémiologie et fondamental pour une bonne représentation de l'ensemble des dépêches. Nous avons analysé les évènements des dépêches en se focalisant sur les détails de déplacement de la population en question. Notre objectif est de construire des historiques des différents cas d'épidémie grâce à une représentation adéquate des phénomènes de déplacement. Pour avoir une telle représentation, notre démarche, telle que le montre la figure 2, repose sur les deux points suivants :

1. **une analyse locale** : cette partie s'intéresse à enrichir le lexique (*Reconnaissance des Entités Nommées*) en identifiant automatiquement des entités nommées utilisées dans le langage des dépêches épidémiologiques (hôpital, organisation, géographie et pathologie).

[1]Integrated Feasibility Experiment for Bio-Security
[2]MITRE Text and Audio Processing

Ensuite, l'analyse des composants des structures d'évènements (personne concernée, lieu, temps et cause) est réalisée grâce à une identification, un étiquetage et une sortie sous forme d'une représentation prédicative logique (*Annotation et représentation des composants*) des composants. Finalement, l'attribution de classes aux verbes du texte (*Typologie des Prédicats*) permet d'avoir une représentation évènementielle sous forme prédicat-arguments où le verbe est le prédicat et les composants sont ses arguments. Il s'agit de reprendre la typologie de Levin utilisée dans *VerbNet* pour les différents prédicats de notre corpus (Levin, 1993) pour cette représentation évènementielle. Cette étape d'analyse est réalisée avec des cascades de transducteurs à nombre d'états fini créées avec le logiciel Nooj[3].

2. **une normalisation** : cette étape a pour rôle d'enrichir certains composants des différentes structures. Nous avons conçu des algorithmes programmés en langage *MetaCard* pour un ensemble de règles de *résolution des pronoms anaphoriques*.

Notre corpus est constitué par des dépêches de l'épidémie du SRAS (*Syndrome Respiratoire Aigu Sévère*), comportant la description des différents moments de la progression de l'épidémie. Pour dégager une méthode d'analyse et de représentation de ces dépêches écrites en langue naturelle, nous avons séléctionné un corpus de 200 dépêches du SRAS. Nous nous sommes intéressés plus spécifiquement à représenter le phénomène de déplacement des individus dans le but d'obtenir une vision de la propagation de l'épidémie. Ce qui nous a amené à sélectionner un sous-corpus d'étude de 100 dépêches, en tenant compte du critère qu'au moins un évènement de déplacement des individus soit présent. Le corpus d'évaluation (21 dépêches) a été choisi selon le même critère à partir des 100 autres dépêches.

3 Analyse locale des dépêches épidémiologiques

Un sous-langage pour les textes scientifiques se caractérise par : un thème scientifique précis, des restrictions lexicales, syntaxiques et sémantiques, des règles de déviance grammaticale (Biber, 1993). Harris considère qu'il est possible de mettre en évidence des classes de mots pour tout sous-langage. Une classe de mots est définie comme un ensemble de mots qui sont acceptables dans les mêmes contextes (Harris, 1968). De telles classes, généralement définies à partir de méthodes distributionnelles, correspondent étroitement aux classes sémantiques qui peuvent être identifiées par un expert du domaine (Grishman *et al.*, 1986). L'analyse des structures évènementielles du sous-langage des dépêches et de ses composants repose sur les méthodes et les théories linguistiques de Harris ainsi que des autres auteurs qui s'en sont inspirés. Nous nous sommes focalisés plus particulièrement sur l'environnement des verbes pour la représentation évènementielle. Avec l'aide des médecins épidémiologistes, nous avons étudié les connaissances associées à chacune des dépêches. Cette étude révèle la présence constante des quatre composants suivants (cf. section 1) autour des actions (verbes) :

– la personne concernée : l'entité complexe qui fait l'objet d'une observation,
– le temps : les différentes expressions temporelles,
– le lieu : les expressions qui font référence à une localisation géographique,
– la cause : les expressions représentant une anomalie médicale, par exemple, un agent pathogène, un virus, une fièvre, une épidémie, *etc.*. De plus, nous nous intéressons aux caracté-

[3]Nooj est un système de traitement automatique de la langue naturelle fondé sur la technologie des transducteurs à nombre d'états fini pouvant être utilisés en cascades.

ristiques tel que le fait d'avoir une maladie contagieuse, d'être en contact avec un malade contagieux et de pouvoir transmettre une maladie contagieuse.

Ces composants reflètent les régularités du sous-langage des dépêches. Ils traduisent ainsi les propriétés conceptuelles du corpus. Leur repérage est utilisé en sortie pour la représentation sous forme de prédicats logiques des différents composants des dépêches. Ces composants se regroupent autour des prédicats verbaux pour former les structures évènementielles. Ce repérage commence par une étape de reconnaissance des principales entités nommées du corpus. La reconnaissance des entités nommées (*REN*) est considérée comme une tâche primaire pour un SEI. De nombreux systèmes ont été développés dans plusieurs langues et dans plusieurs domaines pour reconnaître et catégoriser de telles entités apparaissant dans les textes (Collier *et al.*, 2000; Humphreys *et al.*, 2000; Poibeau, 2003). Ces systèmes se répartissent, selon leurs approches d'analyse, en trois catégories : les systèmes à base de règles écrites à la main, les systèmes à base d'apprentissage et les systèmes mixtes (Sekine & Eriguchi, 2000). Selon ces auteurs, ces trois catégories de systèmes sont équivalentes au niveau de leur performance.

Notre système de REN, pour les dépêches épidémiologiques, est un système à base de règles écrites à la main sous forme de transducteurs appliqués en cascades sur le texte (El-Zant *et al.*, 2006). Les entités identifiées par ces transducteurs sont ajoutées au lexique du système et constituent les entrées de nouveaux dictionnaires après vérification manuelle. Ces dictionnaires ont des priorités plus élevées que celles des dictionnaires standards afin de limiter les ambiguïtés. Nous avons créé quatre grammaires pour le repérage des noms des hôpitaux, des organisations, des pathologies et des localisations géographiques.

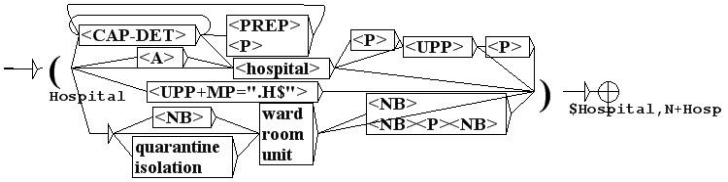

FIG. 3 – Reconnaissance des noms des Hôpitaux.

La figure 3 présente le transducteur de reconnaissance des entités des noms des hôpitaux ou des parties des hôpitaux. Dans ce transducteur, le nœud `<CAP-DET>` signifie que le mot à identifier commence par une lettre en majuscule sans être un déterminant. Le nœud `<UPP+MP=".H$">` signifie que le mot à identifier est entièrement formé de lettres majuscules se terminant par la lettre `H` (sigle d'un nom d'hôpital). La sortie de cette grammaire permet d'étiqueter les noms des hôpitaux, leurs sigles ainsi que divers services d'hôpitaux. Le chemin de transducteur pour l'étiquetage de l'expression *National University Hospital (NUH)* par exemple, est :
`<CAP-DET> <CAP-DET> + <hospital> + <P> + <UPP+MP=".H"$> + <P>`.

Chaque dépêche du corpus d'évaluation est examinée par un épidémiologiste qui donne pour chaque entité le nombre d'occurrence (`NL`). Nous appliquons ensuite les grammaires de REN sur ses dépêches afin de récupérer le nombre d'entités reconnues automatiquement (`NA`). Ce nombre est composé d'un nombre de reconnaissances automatiques correctes (`NAC`) et d'un nombre de reconnaissances automatiques incorrectes ou partielles des entités (`NAI`). L'évaluation globale de notre système pour la reconnaissance des entités nommées est de 97,55% pour la précision ($\mathcal{P} = \frac{NAC}{NAC+NAI}$) et 80,93% ($\mathcal{R} = \frac{NAC}{NL}$) pour le rappel. Nous attribuons ce taux élevé au fait que notre évaluation porte sur un sous-corpus de dépêches (21 dépêches) de l'épidémie SARS et que notre corpus d'étude était beaucoup plus vaste (100 dépêches) et portait sur la même épidémie.

3.1 Annotation et représentation des composants

Le sous-language des dépêches épidémiologiques appartient aux connaissances du domaine décrivant l'évolution des maladies infectieuses. Son lexique, tel qu'il apparaît dans le corpus, regroupe l'ensemble des composants (*personne concernée, temps, lieux* et *cause*, cf. section 3). Pour illustrer cette méthodologie, nous détaillons l'étude et la structure du composant *personne concernée* (cf. figure 4). Une étude syntaxico-sémantique des structures des sous-composants *âge, description, nombre* et *personne* a été nécessaire. Les unités représentant l'*âge* sont formées sémantiquement des deux entités suivantes : un *nombre* (*Age*) et une *unité* d'âge (*AgeUnit*). Ce composant est représenté syntaxiquement soit par un groupe nominal, soit par une séquence verbale construite autour du verbe *age*. La figure 5 illustre l'annotation du composant de l'âge. Pour l'expression *might be aged between 32 and 40 years old* le chemin d'annotation du transducteur est : *PreAge*(might + be+ <age,V>) + *AgeBetween*(<between,PREP>) + *AgeNb*(<NB>) + <CONJ> + *AgeNb*(<NB>) + *Unit*(<year>) + old.

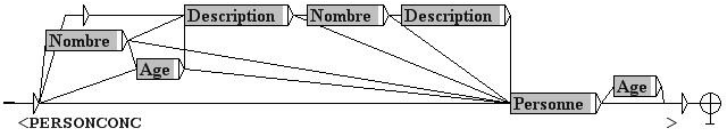

FIG. 4 – Annotation du composant *Personne*.

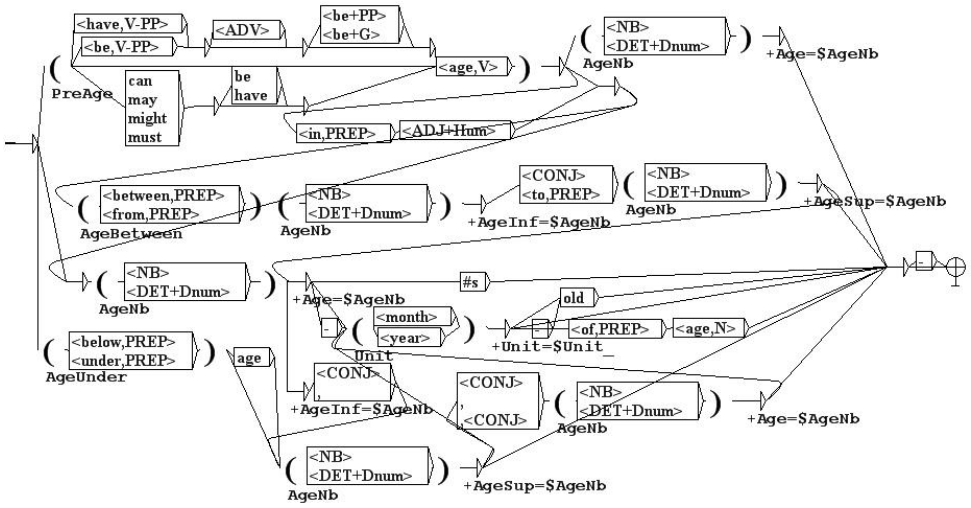

FIG. 5 – L'âge de la personne concernée (composant *Age* de la figure 4).

Le sous-composant *Description* est formé des quatre unités suivantes : *pathologie, cas, localisation et autres descriptions*. La mise en évidence de la pathologie est obtenue par la présence d'un adjectif, d'un nom de pathologie ou de virus qui précède la personne concernée (*a SARS patient, an ill passenger, etc.*). La description locative est le plus souvent un lieu de soins, un pays ou une ville (*two QEH patients, a nurse from Hong Kong, etc.*). La description des différents cas est formée avec des adjectifs représentant le *cas* ou des *numéros* décrivant le classement de ces cas par rapport à la population atteinte d'une épidémie (*new case, the 4th case, etc.*).
Après avoir annoté le composant *personne concernée* avec ces différents sous-composants, un autre ensemble de transducteurs crée en sortie la représentation finale sous forme d'une représentation prédicative logique de ce composant. Ce transducteur possède un niveau supérieur à celui de l'ensemble des transducteurs d'annotation. Son rôle est l'agrégation des composants élémentaires annotés concernant les personnes. La sortie de ce transducteur est un ensemble de prédicats logiques reprenant la sortie de la grammaire d'annotation. Pour l'expression *A 47 year*

old Malaysian who, la sortie d'annotation est : `<PERSONCONC NumPerson="1" Age="47" Unit="year" Description="Malaysian"> A 47 year old Malaysian who </PERSONCONC>` et la représentation logique est donnée par : `<PERSON> NumPerson(p,1) DescripPerson(p,Malaysian) Age(p,47) AgeUnit(p,year) </PERSON>.`
L'évaluation de la représentation des composants est de 97,88% pour la précision et le rappel. Ainsi, pour la représentation des composants, nous avons des valeurs de précision élevées (entre 95,55% et 99,54%). Cette évaluation donne un rappel qui varient entre 97,18% et 99,54%. Notre système obtient un meilleur résultat pour le composant *temps* (97,50% pour la précision et 97,85% pour le rappel). Ceci est dû au fait que ce composant apparaît sous des variétés de forme plus réduites.

3.2 Représentation évènementielle

Une structure évènementielle se compose d'un verbe (*prédicat*) et de ses composants (*arguments*). Chaque argument a un rôle sémantique (ou *rôle thématique*) qui décrit la signification de l'événement indiqué par le verbe (Jackendoff, 1976; Fillmore, 1968). Les rôles sémantiques joués par l'ensemble des arguments par rapport à un verbe particulier entrent dans la structure Prédicat-Arguments de ce verbe. Pour avoir une telle représentation, nous avons adopté pour notre application la typologie de Levin utilisée dans *VerbNet* (Levin, 1993; Kipper *et al.*, 2000). Cette catégorisation se fait avec l'ajout d'un dictionnaire où les verbes les plus significatifs des dépêches se présentent avec un attribut identifiant leurs classes. L'exemple de l'entrée lexicale : `reveal,V+INF+Class=Indicate` indique que le verbe `reveal` est un verbe de la catégorie `Indicate` qui regroupe les verbes fournissant une information indicative (*e.g. explain, prove, reveal, confirm, etc.*). De plus, la représentation de la séquence verbale se fait par l'identification de trois entités : l'auxiliaire `Aux` et/ou l'adverbe `DescripEventTime` s'ils sont présents, le verbe principal `Verb` ainsi que la catégorie sémantique `Class`, la forme affirmative ou négative `Form` et la voix active ou passive `Voice` de ce verbe. La séquence verbale *have been previously examined*, est représentée par la forme logique suivante : `Class(e,Investigate) Verb(e,examine) Aux(e,have been) Form(e,Affirmative) Voice(e,Passive) DescripEventTime(e,previously).`

```
<EVENTSTRUCTURE>
<PERSON> NumPerson(p,3) Person(p,health care workers) </PERSON>
<PREDICAT>Class(e,Show) Verb(e,present) Form(e,Affirmative)
Voice(e,Active)</PREDICAT>
<LOCATION> to(l,hospitals) </LOCATION>
<CAUSE> CausePatho(c,illness) CauseDescripPatho(c,febrile) </CAUSE>
</EVENTSTRUCTURE>
<EVENTSTRUCTURE>
<PERSON>NumPerson(p,2) DescripPerson(p,them)</PERSON>
<PREDICAT>Class(e,Indicate) Verb(e,reveal) Form(e,Affirmative)
Voice(e,Active)</PREDICAT>
<CAUSE> CausePatho(c,pneumonia) </CAUSE>
</EVENTSTRUCTURE>
```

FIG. 6 – Représentation évènementielle (sortie de la phase de l'analyse locale).

Ensuite, la représentation évènementielle se fait grâce à des transducteurs de regroupement des prédicats verbaux et des différents composants qui se présentent dans une même phrase. Ainsi, pour l'exemple suivant : *3 health care workers presented to hospitals with febrile illness. 2 of them revealed signs of pneumonia.*, la représentation évènementielle est donnée par la figure 6 où tout les composants et prédicats sont représentés sous forme prédicative logique.

4 Normalisation

Après avoir extrait les structures évènementielles, le processus de normalisation permet de résoudre certains pronoms anaphoriques afin d'obtenir une représentation évènementielle standardisée. Nous avons mis en œuvre des règles de résolution pour les anaphores pronominales *he, she, they et them*. La règle relative aux pronoms sujet (*He, She et They*) utilise un algorithme similaire pour remplacer l'argument *Personne* représenté par le prédicat logique *Person* ayant la valeur d'un pronom (*Person(p,Pronom)*) par le premier argument *Personne* qui le précède. Ainsi, la règle relative au pronom personnel complément *them* consiste à remplacer le prédicat logique qui contient le pronom *them* `DescripPerson(p,them)` par le prédicat logique `Person(p,_)` de l'argument *personne concernée* qui le précède. Pour l'exemple de la figure 6, l'algorithme de résolution du pronom *them* consiste à remplacer la partie représentant le pronom anaphorique de la figure 6 (`DescripPerson(p,them)`) par `Person(p,health care workers)`.

Pronom	*He*	*She*	*They*	*Them*	Total
Précision	79,59%	70,83%	100%	100%	81,72%

TAB. 1 – Evaluation de la tâche de résolution des pronoms anaphoriques.

L'évaluation de ces règles de résolution des pronoms anaphoriques (tableau 1) est faite en comparant les sorties du programme avec celles fournies par un évaluateur. Les résultats de ce tableau signifient que les pronoms sujet *He, She, they* ont été bien remplacés par l'argument *Personne Concernée* qui les précède et que le pronom complément *Them* a été remplacé par le prédicat logique *Person* de l'argument *Personne Concernée* qui le précède. Notre système donne une précision de 81,72% pour cette résolution que nous estimons comme faible. Ainsi, les valeurs les plus faibles de précision constatées pour les pronoms *He* (P=79,59%) et *She* (P=70,83%) s'expliquent, comme nous le montrons dans l'exemple qui suit, par le fait qu'ils co-réfèrent dans ces cas à des groupe nominaux eux-même anaphoriques. Prenons l'exemple suivant :

> *a 48-year-old welder from Kelantan, who works in Singapore. He had returned to Kelantan on 24 Apr 2003. After falling ill on 26 Apr 2003, he was warded at the Kota Baru hospital. Our investigations revealed that the man did not have any fever when he crossed the Woodlands checkpoint on 24 Apr 2003. Based on our contact tracing, he did not have contact with any SARS patient in Singapore.*

les deux premiers pronoms *he* sont remplacés par l'argument *personne* dont la valeur est *48-year-old welder* et les deux derniers pronoms *he* sont remplacés par l'argument *the man*. Aucune information supplémentaire n'est ajoutée dans ce deuxième remplacement. Pour qu'une telle règle de résolution des cas anaphoriques soit efficace, il faudrait résoudre les cas des groupes nominaux anaphoriques (*the man* par exemple doit être remplacé par *48-year-old welder* pour une bonne résolution des anaphores pronominales).

5 Conclusion

Nous avons mis en évidence le fait que les dépêches que nous traitons sont formées d'un enchaînement de structures évènementielles. Nous décrivons ces structures sous forme d'une représentation argumentale pour laquelle les différents composants *lieu, temps, cause et personne*

que nous avons décrit sont agrégés autour des prédicats verbaux. Dans ce papier, nous avons présenté un système d'extraction d'information permettant d'avoir une représentation évènementielle des dépêches. La normalisation de cette représentation est faite avec des règles de résolution des coréférences. Celles-ci permettent de résoudre les anaphores pronominales (*he, she, they* et *them*). L'évaluation de notre module a été réalisée avec 21 nouvelles dépêches de l'épidémie du SARS. Dans l'avenir, il sera utile de tester notre système d'extraction d'information sur des dépêches concernant d'autres épidémie pour savoir quelle adaptation à apporter aux différentes tâches de ce système afin de renforcer la robustesse de notre SEI.

Références

ABERDEEN J., BURGER J., DAY D., HIRSCHMAN L., ROBINSON P. & VILAIN M. (1995). MITRE : description of the Alembic system used for MUC-6. In *MUC6 '95*, p. 141–155, Morristown, NJ, USA : Association for Computational Linguistics.

BIBER D. (1993). Using register-diversified corpora for general language studies. *Comput. Linguist.*, **19**(2), 219–241.

CHAUDET H. (2004). STEEL : A spatio-temporal extended event language for tracking epidemic spread from outbreak reports. In *KR-MED*, p. 21–30.

COLLIER N., NOBATA C. & TSUJII J. (2000). Extracting the names of genes and gene products with a hidden Markov model. In *Proceedings of the 18th conference on Computational linguistics*, p. 201–207.

DAMIANOS L., WOHLEVER S., PONTE J., WILSON G., REEDER F., MCENTEE T., KOZIEROK R., HIRSCHMAN L. & DAY D. (2002). Real users, real data, real problems : the MiTAP system for monitoring bio events. In *HLT'02*, p. 357–362, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

DAMIANOS L. E., WOHLEVER S., KOZIEROK R. & PONTE J. M. (2003). MiTAP : A Case Study of Integrated Knowledge Discovery Tools. In *HICSS '03*, p.6̃9, Washington, DC, USA : IEEE Computer Society.

DAMIANOS L. E., ZARRELLA G. & HIRSCHMAN L. (2004). *The MiTAP System for Monitoring Reports of Disease Outbreak*. The MITRE Corporation.

EL-ZANT M., ROUX M. & ROYAUTÉ J. (2006). Units' elements of some event structures of outbreaks report. In *InSciT2006*, p. 393–397.

FILLMORE C. J. (1968). The case for case. In E.BACH & R. HARMS, Eds., *Universals in Linguistic Theory*, p. 1–88.

GRISHMAN R. (1995). The NYU system for MUC-6 or where's the syntax ? In *MUC6 '95*, p. 167–175, Morristown, NJ, USA : Association for Computational Linguistics.

GRISHMAN R. (1997). Information Extraction : Techniques and Challenges. In *SCIE '97*, p. 10–27, London, UK : Springer-Verlag.

GRISHMAN R., HIRSCHMAN L. & NHAN N. T. (1986). Discovery procedures for sublanguage selectional patterns : initial experiments. *Comput. Linguist.*, **12**(3), 205–215.

GRISHMAN R., HUTTUNEN S. & YANGARBER R. (2002a). Information extraction for enhanced access to disease outbreak reports. *Journal of Biomedical Informatics*, **35**(4), 236–246.

GRISHMAN R., HUTTUNEN S. & YANGARBER R. (2002b). Real-time event extraction for infectious disease outbreaks. In *HLT '02*, p. 366–369, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

HARRIS Z. (1968). *Mathmatical Structures of Language*. John Wiley Interscience Publishers.

HIRSCHMAN L., CONCEPCION K., DAMIANOS L., DAY D., DELMORE J., FERRO L., GRIFFITH J., HENDERSON J., KURTZ J., MANI I., MARDIS S., MCENTEE T., MILLER K., NUNAN B., PONTE J., REEDER F., WELLNER B., WILSON G. & YEH A. (2001). Integrated Feasibility Experiment for Bio-Security : IFE-Bio a TIDES demonstration. In *HLT '01*, p. 1–5, Morristown, NJ, USA : Association for Computational Linguistics.

HUMPHREYS K., GAIZAUSKAS R. & CUNNINGHAM H. (2000). *LaSIE Technical Specifications*. Rapport interne, Department of Computer Science. University of Sheffield.

HUTTUNEN S., YANGARBER R. & GRISHMAN R. (2002a). Complexity of Event Structure in IE Scenarios. In *Proceedings of the 19th international conference on Computational linguistics*, p. 1–7, Morristown, NJ, USA : Association for Computational Linguistics.

HUTTUNEN S., YANGARBER R. & GRISHMAN R. (2002b). Diversity of Scenarios in Information Extraction. In *Third International Conference On Language Resources And Evaluation*, p. 1443–1450.

JACKENDOFF R. (1976). Toward an Explanatory Semantic Representation. *Linguistic Inquiry*, **7**, 89–150.

KIPPER K., DANG H. T. & PALMER M. (2000). Class-Based Construction of a Verb Lexicon. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, p. 691–696 : AAAI Press / The MIT Press.

LEVIN B. (1993). *English Verb Classes and Alternations*. University of Chicago Press.

POIBEAU T. (2003). *Extraction automatique d'information du texte brut au web sémantique*. Hermès.

SEARS J. A. & CROSS S. E. (2001). The Integrated Feasibility Experiment (IFE) process. In *HLT '01*, p. 1–5, Morristown, NJ, USA : Association for Computational Linguistics.

SEKINE S. & ERIGUCHI Y. (2000). Japanese named entity extraction evaluation - analysis of results. In *COLING*, p. 1106–1110.