Annotation temporelle des événements dans des dépêches épidémiologiques

Yann Guilbaud*, Jean Royauté*

*Laboratoire d'Informatique Fondamentale de Marseille (LIF)
UMR 6166 CNRS - Univ. de Provence - Univ. de la Méditerranée
Parc scientifique et technologique de Luminy
CASE 901, 163 Avenue de Luminy
F-13288 Marseille Cedex 9
yann.guilbaud@lif.univ-mrs.fr, jean.royaute@lif.univ-mrs.fr

Résumé. Nous nous intéressons dans cet article à l'annotation temporelle des événements décrits dans des dépêches issues du site de diffusion d'informations épidémiologiques PROMED MAIL. Après avoir présenté l'intérêt d'un marquage et d'un calcul temporel sur de telles données textuelles pour des épidémiologistes, nous présentons l'expérimentation que nous avons réalisée en utilisant des transducteurs et un calcul exploitant les intervalles d'Allen. Pour le marquage des événements, nous avons été amenés à caractériser le sous-langage de l'épidémiologie tel qu'il apparaît dans les dépêches. Des schémas d'extraction ont ensuite été implémentés sous forme de transducteurs. Le fait qu'il ne soit pas toujours possible de faire référer ces événements à une date précise nous amène à proposer une représentation de la temporalité de ces événements basée sur l'algèbre des intervalles temporels de Allen ainsi que sur ses extensions. Nous utilisons alors les relations décrites entre ces intervalles pour déterminer, à partir des traits temporels explicites, certains traits temporels implicites. Pour faciliter et rendre plus précis le calcul nous avons introduit des scenarii permettant l'extension des relations temporelles explicites en vue de déterminer certaines temporalités implicites. Nous proposons une application qui utilise les annotations obtenues par transducteurs pour réaliser ce traitement. Ce programme, implémenté en Java, complète et calcule les temporalités implicites. Enfin, nous présentons les résultats que nous obtenons à partir de différents corpus réalisés, représentatifs de différentes maladies.

1 Introduction

Les dépêches épidémiologiques sont de courts textes narratifs, de style journalistique, rapportant une succession d'événements connectés décrivant l'évolution d'une épidémie. Ces dépêches constituent une ressource précieuse pour les épidémiologistes chargés du suivi de ces maladies. Cependant sous leur forme actuelle il n'est pas possible d'envisager des traitements automatiques exploitant leur contenu informationnel. Ces textes véhiculent trois types d'informations importantes : une information factuelle qui concerne les événements en eux-mêmes (l'épidémie et ses différentes mani-

festations), une information temporelle qui permet de dater les événements à partir de dates précises, incomplètes ou représentées par des intervalles plus ou moins complexes, et une information spatiale qui situe les différentes localisations géographiques (régions, villes, pays, etc.) caractérisant un événement épidémiologique. Dans cet article, nous présentons le travail de fouille de textes que nous avons réalisé pour l'annotation des événements épidémiologiques, en cherchant à leur fournir à chaque fois une ancre temporelle. La mise en relation de ces événements avec leur temporalité se situe clairement dans le cadre de la fouille de données complexes. En effet, nous ne disposons pas d'une information structurée, exprimable sous une forme tabulaire. Les seules informations qui puissent être repérées de façon régulière sont les dates de chaque dépêche, car elles apparaissent toujours sous la même forme dans un champ clairement identifié. Notre travail porte donc à la fois sur la capture des événements épidémiologiques et des dates de ces événements et sur les calculs temporels qui en découlent.
L'article est structuré de la façon suivante. Dans la seconde section nous présentons le projet EPIDEMIA qui a servi de cadre à ce travail et nous précisons les objectifs que nous nous sommes fixés dans le cadre de ce projet. Nous décrivons dans la section 3 les différentes classes et sous-classes du sous-langage de ces dépêches, les schémas de phrases dans lesquels se trouvent insérés les événements et les différents transducteurs permettant de marquer ces événements et de leur affecter une date. La section 4 est consacrée au traitement temporel des événements. Le fait qu'il ne soit pas toujours possible de faire référence à une date précise nous oblige à définir chaque date d'événement par un intervalle de temps. C'est la raison pour laquelle nous utilisons les intervalles d'Allen. Cette représentation, associée à des scénarios prédéfinis nous permet de situer au mieux une date et de renseigner des dates incomplètes.Nous consacrons la section 5 au calcul des temporalités implicites. Nous traitons les cas d'expressions référentielles et nous illustrons le calcul de Allen en nous intéressant à un exemple concernant le déplacement de personnes. La section 6 est consacrée à une expérimentation permettant de rendre compte des points forts et faiblesses de l'annotation produite.

2 Objectifs et données

Le site PROMED-MAIL (http : \\www.promedmail.org) propose un accès libre à un ensemble de dépêches épidémiologiques. Ces dépêches, recueillies ou écrites par les membres habilités, reprennent des courriels de médecins et de spécialistes, des dépêches d'agences de presse ou de journaux, et des communiqués officiels d'organismes gouvernementaux et non gouvernementaux. Cependant, de part sa nature textuelle, cette information n'est pas sous une forme qui en permette une manipulation automatique. Chaque épidémie, qu'elle touche l'homme ou l'animal, et quel que soit son impact en terme sanitaire, est détaillée et suivie. Ces données sont une source d'information précieuse pour les épidémiologistes qui s'intéressent à des maladies inconnues, ou dont l'incidence et/ou la gravité représente un enjeu de santé publique. C'est la raison pour laquelle, deux équipes du LIF, l'une regroupant des chercheurs en traitement de la langue (équipe CALN - Compréhension Automatique du Langage Naturel) et l'autre dépendant de l'hôpital de la Timone (équipe associée BIM - Biomathématique, Informatique médicale) travaille sur ces données dans le cadre du projet EPIDEMIA. Dans

une perspective de veille sanitaire, l'enjeu est d'être capable d'identifier un faisceau de présomptions et/ou de descriptions cliniques pour prendre les dispositions sanitaires nécessaires le plus tôt possible. Par exemple, pour l'étude des épidémies, il est important d'être capable de mettre en corrélation des événements distincts mais reliés au même phénomène épidémiologique : ainsi, il a été ainsi démontré que le virus Ebola atteignait d'abord les primates des forêts équatoriales d'Afrique avant d'atteindre les populations humaines. Or, il est parfois délicat de mettre en corrélation des événements pourtant liés, et ce du fait de la dissémination de l'information. L'ambition du projet EPIDEMIA est de construire une modélisation du contenu de ces dépêches sous la forme d'une représentation de l'évolution des caractéristiques de l'histoire de chaque épidémie à partir des informations spatio-temporellement localisées qui sont contenues dans les observations fragmentaires que sont les dépêches.
Le travail expérimental que nous présentons ici poursuit un objectif triple plus limité : identifier dans les dépêches les modes d'expression de la temporalité des événements, puis annoter ces expressions, et enfin déterminer certaines informations temporelles implicites. Le terme événement désigne habituellement une description d'une modification du monde. Nous le définissons plus formellement comme une description statique ou dynamique des éléments du monde. Une modification du monde est l'énoncé de l'altération de l'état du monde, état précédemment décrit ou implicitement déduit. Une description statique est un énoncé qui présente, sans que celles-ci ne soient modifiées, les propriétés du monde ou de ses éléments. Désigner par événement ces expressions peut se justifier d'une part par le fait que l'énoncé d'une propriété modifie le monde, ou du moins modifie la description de celui-ci, et d'autre part par le fait que ces propriétés sont, dans le cadre de cette étude et plus généralement en sciences, des hypothèses amenées à être modifiées.
Nous cherchons donc à assigner des propriétés temporelles à certains événements décrits dans les dépêches épidémiologiques PROMED (cf. section 3). On désigne habituellement par propriétés temporelles - ou temporalité - les informations de date, de durée et/ou de fréquence. Nous verrons dans les sections qui suivent comment nous représentons ces informations et ensuite quel type d'informations temporelles nous avons valorisé (cf. section 4.1).

3 Sous-langages de l'épidémiologie

Afin d'établir le marquage des événements, il nous a fallu réaliser l'étude des différentes classes et sous-classes du sous-langage de l'épidémiologie. Cela nous a obligé à préciser quels événements vont être marqués, puis à étudier les différents modes d'expression de ces événements dans les dépêches.

3.1 Terminologie de l'épidémiologie : classes et sous-classes

Nous avons été confronté lors de notre étude à différentes difficultés propres au domaine, certaines d'entre-elles étant liées aux principes de fonctionnement du site PROMED. Une première difficulté a été le manque d'homogénéité des classes. En effet, de nombreux termes sont présents dans plusieurs classes. Les symptômes d'une

maladie peuvent désigner, par exemple, la maladie elle-même : ainsi une fièvre peut être considérée, sémantiquement, comme une maladie, alors qu'il s'agit en fait d'un symptôme qui est décrit pour de nombreux tableaux cliniques. D'autres termes, comme ceux désignant les personnels techniques, peuvent être, sémantiquement, à double voire triple emploi : "doctor" peut par exemple désigner le personnel soignant, une personne faisant une déclaration ou encore une victime d'une épidémie. Seul le contexte nous permet de distinguer ces occurrences. Ce manque d'homogénéité est dû au domaine épidémiologique. Une autre difficulté à signaler est le large champ lexical utilisé dans ces dépêches. Le domaine traité implique notamment une connaissance de nombreux termes de géographie, et des correspondances entre ceux-ci. De même, les noms des différentes personnes exprimant des informations (Ministres, porte-parole, responsables etc.) doivent être relevés et identifiés à leur poste. L'éventail des symptômes pouvant toucher une personne justifierait à lui seul une étude complète, et comme pour le reste du vocabulaire, nous avons procédé à un relevé en contexte d'une partie de ces termes afin de les intégrer à nos transducteurs. Par ailleurs, du fait du fonctionnement du site PROMEDMAIL (pluralité des auteurs et des sources), les dépêches ont une structure variée, un vocabulaire plus ou moins spécialisé, de nombreuses tournures différentes pour exprimer une même information. Cela peut s'avérer problématique, car en multipliant les tournures de phrases, on augmente le risque que certaines d'entre-elles se retrouvent dans un contexte semblable, mais dans des sens différents. Nous avons pu contenir ce problème en traitant non pas les expressions par parties, mais par blocs d'expressions : au lieu de repérer individuellement des séquences, nous relevons des séquences de séquences. Nous décrivons dans les tableaux suivants les principales classes et sous classes d'expressions rencontrées (cf. TAB .1). Les expressions en italique désignent des termes appartenant à la classe précisée. La classe TEMPORALITE rassemble les termes ou expressions utilisés dans notre corpus pour décrire les caractéristiques temporelles des événements. On distingue ensuite les expressions calendaires ou horaires, au sens strict, des constructions grammaticales dans lesquelles on les rencontre. De même, au sein de la classe GEOGRAPHIE, on distingue les noms de lieux des constructions exprimant une localisation statique ou dynamique. La classe EPIDEMIE contient les termes ou expressions du domaine de l'épidémiologie, ce qui inclus des termes de médecine, de biologie, de chimie et de pharmacie. La classe ENTITES contient les termes ou expressions désignant des personnes physiques ou morales. La classe VERBES rassemble les verbes intervenant dans les compositions entre éléments des classes précédentes. Ces verbes sont relevés dans le contexte des éléments déjà classés.

3.2 Schémas de phrases et événements

A partir des définitions de classes et sous-classes précédentes, nous utilisons la composition d'éléments de ces différentes classes pour décrire des événements de nature épidémiologique. Ainsi, la composition d'un élément de la classe EPIDEMIE, suivi d'un élément de la classe VRB-EPI, suivi d'un élément de la classe ENTITES, dénote un événement épidémiologique. Nous décrivons ainsi quatre classes d'événements.

Classe	Sous-classe	Description	Exemple
TEMPORALITE		Temporalités	
	DATE-HEURE	Date calendaire ou horaire	*Monday, today, last month, ...*
	TEMP-PONC	Date ponctuelle	*On DATE-HEURE, As of DATE-HEURE, ...*
	TEMP-INTER	Période de temps	*Between DATE-HEURE and DATE-HEURE, ...*
GEOGRAPHIE		Localisation spatiale	
	GEO-NAMES	Nom de lieu	*Congo, Paris, zone, ...*
	GEO-STATIC	Localisation statique	*In GEO-NAMES, at GEO-NAMES, ...*
	GEO-DYNAMIQ	Localisation dynamique	*To GEO-NAMES, from GEO-NAMES*
EPIDEMIE		Maladie, épidémie, vecteur de maladie ou action sanitaire	
	EPI-MALADIE	Maladie	*Influenza, SARS, ...*
	EPI-VECT	Vecteur ou un hôte de maladie	*Hantavirus, micoplasma, body fluids, ...*
	EPI-SYMPTOMS	Symptôme, constante biologique ou résultat	*Aches, fever, 39.4° C, X-ray abnormalities, ...*
	EPI-ACT	Action sanitaire ou thérapeutique	*Disease control measure, mechanical ventilation, ...*
ENTITES		Entité, personne physique ou morale	
	ENTITIES-NAMES	Personne morale	*World Health Organization, ONU, ...*
	PER-NAMES	Nom propre	
	PER-FAMILY	Membre d'un groupe social	*Wife, son, friend, ...*
	PER-TECHS	Personnel technique	*Nurse, doctor, ...*
	PER-TEAMS	Groupe d'individus	*Hospital staff, a team of PER-TECHS, ...*
	PER-SPK	Personne s'exprimant	*Officials, a ENTITIES-NAMES spokesman, ...*
	PER-TRAVEL	Personne ou groupe se déplaçant	*Traveller, airline crew, etc.*
	PER-VICTIMS	Victime, cas ou patient	*Case, deaths, contacts, etc.*
VERBES		Verbes	
	VRB-COM	Communication ou recommandation	*To say, to inform, to appeal, ...*
	VRB-ACT-SAN	Action sanitaire ou thérapeutique	*To isolate, to diagnose, ...*
	VRB-EPI	Epidémie, maladie ou symptôme	*To kill, to contract, to suffer, ...*
	VRB-DEPLACE	Déplacements d'individus	*To leave, to fly, to visit, ...*

TAB. 1 – Classes et sous-classes du sous-langage de l'épidémiologie

3.2.1 Evénements de nature épidémiologique

Apparitions, Contagions et Victimes : nous regroupons dans cette catégorie les événements qui définissent l'apparition d'une maladie ou d'une épidémie, ceux qui font état d'une propagation de celle-ci ou qui décrivent ses modes de transmission et enfin les descriptions du nombre et du type d'individus touchés. Exemple : *Ebola killed 5 people, Villagers have contracted Avian Influenza, etc.*
Action sanitaire et thérapeutique : nous désignons sous cette appellation les événements faisant état de mesures sanitaires prises par les autorités ainsi que des démarches thérapeutiques mises en oeuvre dans le traitement des maladies ou épidémies. Exemple : *Authorities have isolated the area, Infected people have been treated with antivirals, etc.*
Description de cas : nous appelons description de cas, les événements qui caractérisent les symptômes d'une maladie ou d'une épidémie, ainsi que ceux qui décrivent les cas cliniques, les tableaux diagnostiques et leur évolution. Exemple : *High fever was diagnosed, M. X complained about headaches, etc.*
Déplacements d'individus : ces informations étant très pertinentes en épidémiologie, nous regroupons sous cette catégorie les événements qui font état du déplacement de personnes - et qui par conséquent renseignent sur les lieux d'exposition à une maladie ou une épidémie, sur la date de cette exposition etc. Exemple : *Inhabitants are leaving the area, Villagers are arriving in town, etc.*

Nous utilisons pour identifier et annoter les événements des transducteurs à nombre fini d'état (Finite State Transducers ou FST) créés grâce au logiciel INTEX. Cette approche est proposée dans de nombreuses études sur l'extraction d'informations en langage naturel [Schilder, 1999] [Schilder, 2001] [Grishman *et al.*, 2002]. Il s'agit de graphes dont les noeuds (ou étiquettes) sont les termes à reconnaître et dont les arêtes définissent les séquences de termes acceptables. Grâce à ces transducteurs nous pouvons insérer ou substituer du texte à celui relevé. D'autres approches d'annotations temporelles ont été expérimentées par ailleurs, notamment dans [Filatova, 2001], où les événements sont repérés dans des arbres syntaxiques, ou encore dans [Mani et Wilson, 2000] où, est réalisée une annotation temporelle se basant sur des expressions temporelles et utilisant un étiquetage en parties de discours (POS Tagging). L'utilisation des transducteurs plutôt que d'autres outils se justifie par le fait que la complexité de la réalisation de notre annotation est linéaire et uniquement fonction du nombre de mots de la séquence à traiter, mais aussi par le fait que les transducteurs sont aisément mis à jour ou modifiés. Par ailleurs, le choix de procéder à un étiquetage plus fin, qui traite un phrase entière plutôt que d'identifier des éléments épars d'informations, permet de pouvoir ré-exploiter les annotations XML réalisées, les enrichir ou les modifier dans d'autres applications. De plus, la conservation, dans le document annoté avec un balisage de type XML, de la quasi totalité du texte original rend cette annotation transparente et vérifiable. Enfin, nous avons choisi un balisage de type XML afin de garantir un accés facile et personnalisable aux informations annotées. XML est en effet conçu pour assurer la traçabilité de l'information, via une interface ou une API conçue en fonction des besoins spécifiques d'un utilisateur. Toutefois, la conception d'une telle interface de démonstration n'a pas encore été réalisée pour ce projet.

3.3 Transducteurs d'annotation et de normalisation

Nous avons été amenés à concevoir plus de 600 transducteurs dans le cadre de ce travail. Si certains sont redondants (transducteurs appropriés à différents traitements d'une même séquence), cela représente plus de 200 ensembles de séquences identifiables individuellement (plus de 200 types d'expressions). Pour identifier les différents modes d'expression d'un même concept, ou d'une même catégorie de concepts, nous devons être capables d'identifier une séquence d'items. Grâce à la possibilité d'étiqueter un noeud par un transducteur, nous pouvons reconnaître des séquences de séquences. Par exemple, la suite (Sujet+Verbe+Compléments) peut être identifiée en reconnaissant un Sujet par un transducteur dédié, suivi d'un verbe (identifié par un autre transducteur), lui-même suivi de compléments (identifiés grâce à des transducteurs dédiés). De plus cette utilisation permet un calcul plus efficace lors de l'identification de mots et de séquences. En effet, la complexité de l'identification n'est dépendante que de la taille de l'unité à traiter. L'utilisation que nous faisons des transducteurs (cf. fi-

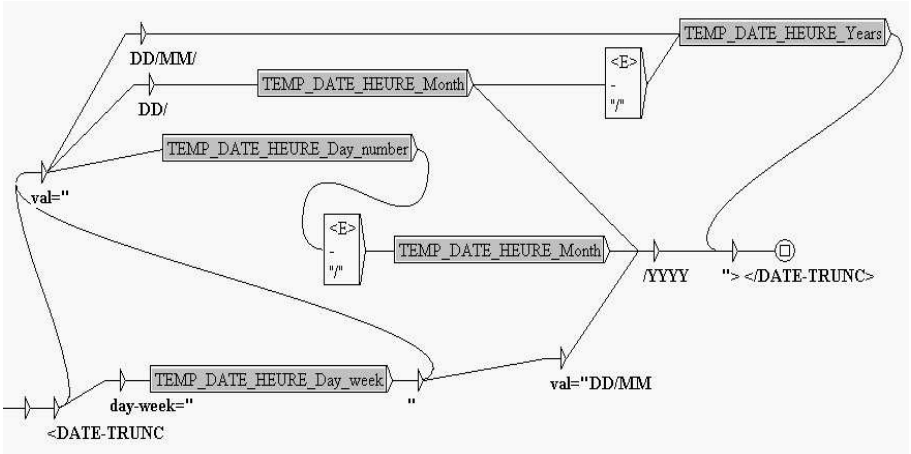

FIG. 1 – Transducteur d'annotation des dates incomplètes

gure 1) consiste à capturer et encapsuler dans des balises XML les informations de date présentes dans le texte. Nous utilisons aussi des transducteurs pour substituer la séquence reconnue par une forme normalisée de cette séquence, mais avec le risque de perdre définitivement de l'information en cas d'erreur. Cette possibilité est toutefois intéressante pour la normalisation d'un texte (c'est à dire la mise sous une forme canonique des différentes variantes possibles d'une expression). Si une séquence peut être reconnue en concurrence par plusieurs transducteurs, l'opération d'annotation est impossible. Cela a deux conséquences. Tout d'abord, nous devons réaliser séquentiellement les opérations d'annotations concurrentes, en se basant sur une relation d'ordre (par exemple reconnaître une date complète doit être réalisé avant de chercher à reconnaître une liste de dates incomplètes, et cette deuxième opération ne doit pas opérer sur les parties de texte déjà étiquetées). Par ailleurs, pour les transducteurs concurrents ne

pouvant être classés, nous devons être capables d'utiliser le contexte pour lever les ambiguïtés. Cela peut se faire, soit en se basant sur des balises déjà insérées, soit en reconnaissant des blocs de séquences. Après différents essais, nous avons été amenés à choisir la deuxième solution, qui, même si elle est considérablement plus coûteuse en ressources, est beaucoup plus fiable, car pour la première, le choix de l'ancre pour démarrer le balisage peut s'avérer impossible. Pour pouvoir accéder plus aisément aux informations contenues dans les dépêches, nous insérons des balises de type XML dans le texte traité. Le schéma d'annotation à retenir est fortement dépendant du domaine et de l'expertise humaine, nous avons donc retenu un schéma déjà existant en le simplifiant [Grishman *et al.*, 2002], mais ces schémas peuvent être appliqués à des textes plus généraux, comme l'ont montré [Filatova, 2001] et [Setzer et Gaizauskas, 2000]. A chaque séquence identifiée, nous faisons correspondre une balise spécifique. Grâce à la possibilité d'imbriquer des transducteurs en les appelant les uns au sein des autres, nous obtenons en sortie un balisage imbriqué. Le fait de normaliser en partie les textes dans un prétraitement permet une structuration assez rigide de nos balises. Certaines balises comme <*DATE-NORM*> ou <*DATE-TRUNC*>, qui sont des marqueurs de date, sont complétées avec une valeur extraite du texte, ou encore avec un marqueur que nous substituons par sa valeur lors du traitement de nos balises. Le choix d'opérer un traitement séquentiel nous fournit un document quasiment structuré. En effet, l'encapsulation des balises est garanti par le protocole choisi. Les balises principales que

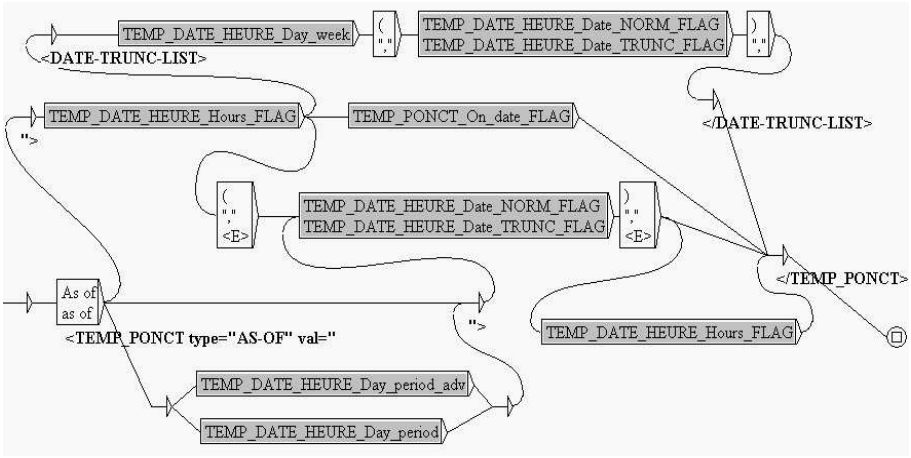

FIG. 2 – Transducteur d'annotation des compléments introduits par "As of" ou "As at"

nous utilisons sont :
- pour les dates, <*DATE-NORM*> (complètes) et <*DATE-TRUNC*> (incomplètes),
- pour les expressions de date, <*TEMP-PONCT*> (ponctuelle) et <*TEMP-INTER*> (période),
- pour les expressions d'événement, <*EVEN*>,
- pour les localisations, <*GEO-STATIC...*> (statique) et <*GEO-DYNAMIQ*> (dyna-

mique).
La plupart de ces balises ont été décrites dans la section 3 concernant le sous-langage de l'épidémiologie.
La figure 3, montre comment l'information de la phrase suivante :
As of Tue 11 Dec 2001, WHO has received reports of 12 suspected cases, including 10 deaths from haemorrhagic fever.
est capturée et encapsulée dans des balises XML. Cette phrase est reconnue par un transducteur ad hoc comme l'expression d'un événement de type REPORT-BILAN. Si l'on s'intéresse uniquement au marquage du temps de ce transducteur, la séquence *As of Tue 11 Dec 2001* qui est un circonstant de temps est identifiée ici par appel du transducteur 2. Une dernière étape de transduction est nécessaire pour normaliser la date et ajouter cette information dans les attributs *dayweek* et *val* de la balise <*DATE-NORM*>. L'annotation finale de cette phrase se trouve dans la figure 3.

```
<EVEN type="REPORT-BILAN">
  <TEMP-PONCT type="AS-OF" val="As of">
     <DATE-NORM dayweek="TUE" val="11/12/2001"/>
  </TEMP-PONCT>
  ,
  <ORG-WHO>WHO</ORG-WHO>
  <VRB val="RECEIVE" />
  <EPI-MANIF-OF-MALADIES>
     <EPI-MANIF-COMBOS>
        <EPI-MANIF-STAT>reports of 12 suspected cases</EPI-MANIF-
STAT>
        , including
        <EPI-MANIF-STAT>10 deaths</EPI-MANIF-STAT>
     </EPI-MANIF-COMBOS>
     from
     <EPI-MALADIES>haemorrhagic fever</EPI-MALADIES>
   </EPI-MANIF-OF-MALADIES>
  .
</EVEN>
```

FIG. 3 – Annotation d'un événement de type REPORT-BILAN

4 La temporalité et les relations temporelles entre événements

4.1 Représentation par intervalle : intervalles de temps de Allen

La représentation la plus immédiate de la temporalité consiste à placer les événements sur un axe des temps, ou axe des dates. Depuis les travaux de James F. Allen [Allen, 1984], nous disposons d'un formalisme de description de la temporalité par intervalle. Un événement est ainsi décrit temporellement par l'intervalle de temps, borné, qui contient son début et sa fin. Il est important de noter que l'intervalle choisi contient les bornes réelles de l'événement, mais qu'en aucun cas nous ne pouvons avoir la prétention de déterminer absolument les dates de début et de fin. La datation absolue reste une

chimère, ne serait-ce que par la granularité de notre conception du temps : même la date la plus complète, associée à une heure aussi précise que possible n'est qu'une approximation. De plus, de part la forme des dépêches et de part l'incertitude quant aux événements décrits, même les événements ponctuels sont avantageusement décrits par des intervalles. En effet, quelle que soit la granularité temporelle choisie (l'échelle de temps choisie), une date même ponctuelle n'est définie qu'à cette échelle de temps prés (ainsi, si l'échelle est la journée, la précision est à 24 heures prés). Par exemple, l'expression "On Mon 25 August" désigne l'intervalle de temps pour le lundi 25 Août compris entre zéro heure et minuit. La représentation par intervalle reflète cette caractéristique. Par suite, dans la description des intervalles de temps compris entre deux dates, nous utiliserons deux intervalles de Allen, un pour chaque date. Cette représentation est, par ailleurs, proposée dans l'annotation de la temporalité dans des dépêches [Setzer et Gaizauskas, 2000], dans les historiques thérapeutiques [Bouaud, 2004] ou dans les rapports d'épidémie [Grishman *et al.*, 2002]. Si un événement est une ins-

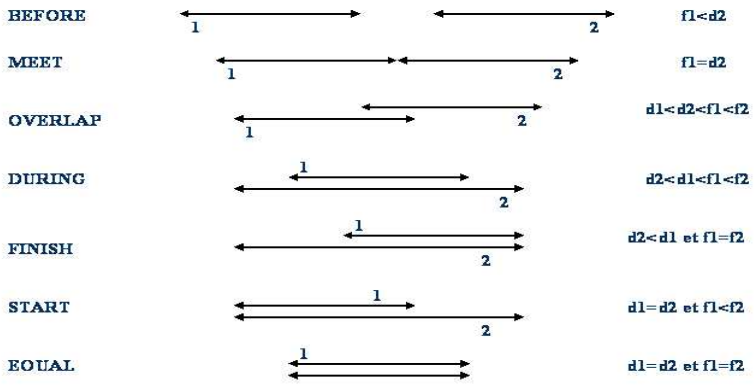

FIG. 4 – Relations de Allen

tance d'un phénomène périodique, il est représenté par plusieurs intervalles de temps, un pour chaque occurrence du phénomène. Le formalisme choisi est donc applicable à la périodicité, et il est techniquement possible de déterminer la périodicité à partir des différents intervalles de Allen de chaque instance. Toutefois, être capable de réaliser ce traitement implique d'identifier les instances d'un phénomène périodique, ce qui, dans le cadre du traitement automatique que nous présentons ici, dépasse de loin notre propos. Par la suite, nous considérerons que chaque instance d'un phénomène périodique est un événement unique et indépendant. La durée d'un phénomène peut être aisément approximée à partir de l'intervalle de temps qui le décrit. Nous pouvons donc choisir les intervalles de Allen comme représentation des temporalités que nous voulons traiter. Nous avons vu précédemment (cf. section 2) que l'on pouvait attribuer, au minimum, une borne aux descriptions que nous voulons dater. En tout état de cause, nous pouvons fixer l'autre borne de l'intervalle arbitrairement, en prenant comme date butoir

inférieure la date de parution de la première dépêche parue sur PROMEDMAIL (19 Août 1994), et comme date butoir supérieure celle de la dépêche traitée. Entre deux intervalles de Allen il existe 13 relations possibles : 6 relations orientées et leurs contraposées, et la relation d'égalité (cf. FIG. 4). Ces relations sont définies par comparaison entre les bornes de ces intervalles. Il est important de noter qu'il n'existe pas d'autres relations possibles.

4.2 Extensions des travaux de Allen

Grâce à la représentation par intervalles et grâce aux relations décrites, il est possible de représenter par un graphe orienté les événements et leurs relations temporelles [Allen, 1983]. Les noeuds du graphe sont les événements (qui sont associés, chacun, à au moins un intervalle de Allen) et deux événements sont reliés par une arête labellisée avec la relation entre leurs intervalles de temps. Il est possible d'inférer une relation entre deux événements d'après leurs relations à un même troisième, en utilisant les propriétés de transitivité de certaines relations comme la relation AFTER ou BEFORE. De plus, comme il est toujours possible de déterminer une relation entre deux intervalles connus - en comparant leurs bornes - le graphe décrit ci-dessus est complet. On peut donc, à partir des intervalles de Allen déterminer toutes les relations temporelles entre des événements datés. Mais, si on sait que l'on peut donner un intervalle de Allen pour tout événement que nous traitons, nous ne pouvons garantir la précision de cet intervalle. En l'absence de marqueurs de temps, le mieux que nous puissions faire est de donner deux bornes arbitraires (cf. section 4.1). Or, notre connaissance du domaine permet en réalité une plus grande précision. En effet, il existe des relations temporelles implicites entre événements, en particulier quand ceux-ci constituent des séquences. Par exemple, les différentes phases d'un voyage sont temporellement liées : il est possible de déterminer la temporalité d'un vol en l'extrapolant à partir de celle des dates de départ et d'arrivée de celui-ci. De même, les événements épidémiologiques, en particulier les descriptions de cas cliniques et de l'évolution des symptômes, appartiennent à une séquence dont la connaissance nous permet d'extrapoler les temporalités de ses constituants. Ainsi, il est possible, grâce à notre connaissance du domaine d'inférer des caractéristiques temporelles à partir des relations temporelles implicites entre événements. Nous avons vu précédemment qu'il existait des événements prenant place entre deux dates. Nous avons convenu de les représenter par deux intervalles de Allen (un pour chaque date). Or, nous ne pouvons plus, dès lors, nous contenter des 13 relations décrites ci-dessus : nous ne serions capables que de comparer borne par borne ce type d'intervalle. Il nous faut donc définir 168 nouvelles relations, dépendant des relations entre ces bornes ; ces nouvelles relations sont en fait un couple de relations classiques, un membre de la paire décrivant la relation pour les bornes inférieures, l'autre pour les bornes supérieures, d'où 13*13=169 relations, moins l'égalité, ce qui fait un total de 168 relations. Il est possible de comparer ce type d'intervalle avec un intervalle simple en considérant celui-ci comme un intervalle double. Par exemple, l'intervalle simple (01/01/05 ;31/12/05) est équivalent à l'intervalle double ((01/01/05 ;01/01/05) ;(31/12/05 ;31/12/05)). Ainsi, nous pouvons déterminer le graphe temporel complet des événements d'une dépêche.

5 Calcul des temporalités implicites

5.1 Scénarii de calcul des temporalités

Le traitement commence par un parsage du document XML, et le prétraitement des dates repérées. Nous commençons par convertir les temporalités au format "Allen", ce qui fait que toute date est exprimée par deux valeurs : sa borne supérieure et sa borne inférieure (cf. section 4.1). Cela se fait, lors de la création de l'arborescence de notre document, en dupliquant les noeuds de type balises de dates (balises de type *<DATE-NORM>* et *<DATE-TRUNC>*). Les listes de dates incomplètes, encapsulées par des balises *<DATE-TRUNC-LISTE>*, sont prétraitées en propageant de proche en proche les valeurs connues pour renseigner les valeurs manquantes dans ses composantes. Quand nous éditons un noeud de type *<DATE-TRUNC>*, nous réexaminons son contenu avant de le réécrire. S'il correspond à une date complète, nous recopions les valeurs dans un nouvel élément de type *<DATE-NORM>*. De même, les listes de dates incomplètes sont réévaluées et traitées en conséquence.

5.1.1 Résolution de références

Les références temporelles que nous avons rencontrées sont de deux types. Le premier type, le recours à des anaphores comme "today", est traité comme une date incomplète. Dans la plupart des cas, ces anaphores sont explicitées dans le texte par une indication de date entre parenthèses. Par exemple : l'expression *"yesterday (today, 22 May 2001)"*, désigne, par un couple composé d'une date incomplète ("yesterday" que nous normalisons comme un jour de la semaine, avec le marqueur *PRVD*) et d'une date complète ("today, 22 May 2001" marqué par *CURD* comme un jour de semaine), un seul jour : le 21 May 2001. Les couples de date de ce type sont fusionnés. Pour les autres cas, nous précisons par un marqueur approprié la valeur attendue, qui est renseignée par l'application : par exemple "*CURD*" pour la résolution de l'anaphore "today" qui se base sur la valeur de la date de la dépêche (marquée par la balise *<DepecheDATE...>*, insérée dans l'entête d'une dépêche). Le deuxième type d'anaphores est la référence à un événement pour ancrer temporellement un autre événement. Dans ce cas, l'annotation nous fournit le type de relation temporelle décrite (définie par l'adverbe, la préposition ou le pronom relatif qui les relie), et si l'événement référencé est aisément identifiable (dans notre cas, si un événement du voisinage est du même type), nous pouvons déterminer une temporalité [Lascarides et Oberlander, 1993] [Schilder, 1999].

5.1.2 Scénario de description de symptômes et de leur évolution

Dans [Bouaud, 2004], il est décrit une méthode d'inférence des données temporelles quant à l'historique thérapeutique d'un patient. Nous appliquons cette méthode aux descriptions de l'évolution des symptômes. Nous avons rencontré deux constructions que nous souhaitions exploiter pour déterminer la temporalité de ces événements. Nous avons tout d'abord rencontré une description phase par phase des symptômes (phases successives ou se chevauchant, i.e. relations MEET ou OVERLAP). Nous rencontrons aussi une description des évolutions par l'utilisation de propositions relatives. Nous avons convenu que pour une même catégorie de symptôme, la succession des événements

était de type MEET. L'apparition d'un nouveau symptôme, même s'il ne remplace pas un symptôme précédent, crée un nouveau tableau clinique qui succède au suivant. Pour simplifier les traitements, nous procédons de même pour les symptômes qui ne sont pas enracinés : nous considérons qu'ils sont en relation de type MEET avec le symptôme précédant même si celui-ci est de type différent. A partir de ces relations et des temporalités explicitées de certains événements (nous connaissons au moins, dans tous les cas, les dates des premiers symptômes et la date de décès le cas échéant, ou la date de guérison), nous déterminons les temporalités des descriptions de symptômes ou de leur évolution qui ne sont pas datées. Nous ne recherchons ces informations que dans les événements à proximité immédiate, en les propageant de proche en proche, ce qui nous permet de limiter la complexité de ce calcul à un $o(n^2)$ où n est le nombre d'évènements de la dépêche.

5.1.3 Scénario de déplacement de personne(s)

Les descriptions de déplacements d'objets obéissent à un schéma fixe. En étendant la représentation usuelle de l'aspect de la localisation spatiale, présentée dans [Borillo, 1998], on peut décomposer un déplacement en cinq phases. La première est la phase où l'objet quitte le lieu d'origine. La deuxième est celle où l'objet entre dans le lieu de transit. La troisième signale quand l'objet est dans le lieu de transit. La quatrième est la phase où l'objet quitte le lieu de transit. Enfin, la cinquième phase est la phase où l'objet entre dans le lieu de destination. Ces cinq phases se succèdent temporellement, avec une relation de Allen de type "MEET". La connaissance de la temporalité d'une partie de ces phases permet donc d'inférer la temporalité de certaines des autres phases. Dans notre cas, nous ne repérons les phases que si elles sont décrites successivement, qu'elles soient datées ou non. En effet, l'identification de co-référence à un même événement, en l'occurrence un voyage, sont déterminable dans une suite de propositions [Danlos, 1999]. Mais notre propos n'est pas d'identifier ces phases d'après leur expression dans des parties distinctes d'un texte. Pour réaliser le traitement, nous utilisons la liste des événements, et quand nous rencontrons un événement de type VOYAGE (*<EVEN type="VOYAGE"...>*) dont une date est incomplète, nous recherchons l'information manquante dans son entourage immédiat. Lors du traitement de séquences comme *he travelled from Berlin to New-York via Paris on Monday, 11 March 2003*, nous insérons des balises pour deux événements, un événement (*<EVEN...>*) de type VOYAGE "de Berlin à Paris", et un autre "de Paris à New-York" "le Lundi 11 Mars 2003". Or, dans un traitement purement linguistique, l'attribut de date correspondant à ce voyage n'est rattaché qu'à l'un des deux (le plus proche), alors que sémantiquement, il s'applique aux deux composantes. Pour cette raison, nous cherchons une information manquante aussi bien avant qu'après l'événement concerné. Les informations manquantes en fin de traitement sont renseignées avec les valeurs extrêmes valides pour le champ considéré. Par exemple, pour la date [DD/MM/2003;DD/MM/2003], 01 et 12 sont substitués pour les valeurs de mois. Le jour est lui renseigné avec les valeurs 01 et 31, pour obtenir l'intervalle de Allen : [01/01/2003;31/12/2003]. Si nous nous intéressons à la séquence suivante :

He travelled to New-York between 22 Mar and 27 Mar 2004. He embarked in Frankfurt.

He landed at 10 PM on Monday 22. He flew back on Saturday 27 Mar. When the flight landed, he went to the hospital.

nous pouvons, ainsi que le montre la figure 5, la décomposer en 6 événements de type VOYAGE. A chacun de ces événements est associé un verbe, de balise VRB, avec en attribut la forme infinitive de ce verbe. Ces événements sont tous marqués temporellement dans le champ ALLEN avec des intervalles de dates complets ou incomplets. Le premier événement (TRAVEL) est introduit avec le verbe d'attribut TRAVEL. La préposition *between* qui suit, définit une période de temps marqué par deux intervalles d'Allen. Cette information est signalée par la balise TEMP-INTER d'attribut BETWEEN. Ces intervalles délimitent ici les bornes temporelles extrêmes du scénario. Le second événement (EMBARK) caractérise la borne inférieure du scénario, exprimé par le verbe d'attribut EMBARK. Le troisième événement (LAND1) caractérise la phase 4 d'un voyage, exprimée par le verbe d'attribut LAND. Cet événement est positionné temporellement par une date incomplète, signalée avec la balise TEMP-PONCT, grâce à la préposition *at* qui donne ici une précision horaire (*at 10 PM*) à une date déjà citée. Le quatrième événement (FLY) marque ponctuellement une autre date qui correspond à la borne supérieure du scénario. Il est introduit par le verbe d'attribut FLY. Cet événement est ancré temporellement par une date incomplète. Le cinquième événement (GO), centré autour du verbe d'attribut GO, en tant que dernier événement, signale la fin du scénario VOYAGE. Il correspond à la phase 5 du voyage. Il est de nature particulière dans la mesure où il encapsule le sixième événement (LAND2). Dans GO, la balise TEMP-REL-WHEN, déclanchée par la conjonction *when*, signale la nature relationnelle de cet événement qui induit une précédence courte de LAND2 par rapport à GO. Cet événement GO, construit autour du verbe d'attribut GO, décrit l'épilogue du scénario VOYAGE, qui se termine par l'acheminement à l'hôpital du voyageur. LAND2 correspond à la phase 4 d'un voyage et explicite la temporalité de GO. L'examen de

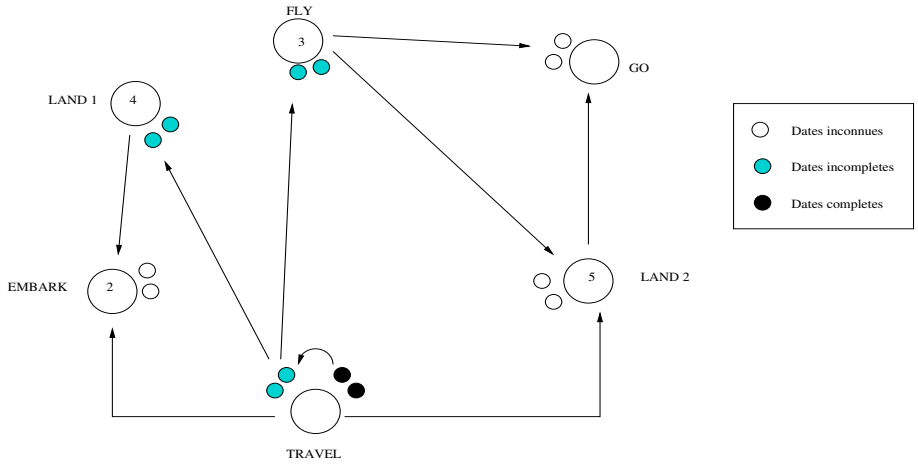

FIG. 5 – Graphe de propagation des informations temporelles

la figure 5 montre que l'ensemble des champs date est incomplet. Les arêtes du graphe de cette figure illustrent la propagation de l'information temporelle d'un événement à l'autre. L'événement TRAVEL, où les bornes supérieures de l'intervalle BETWEEN sont complètement renseignées, permet de propager l'information d'année aux bornes inférieures de cet événement. En effet, le scénario de voyage implique des déplacements courts, qui n'excèdent pas une année dans le cadre des dépêches épidémiologiques, il est donc possible de propager aux événements LAND1 et FLY les informations manquantes d'année. Pour EMBARK, où aucune information de date n'est présente, nous allons pouvoir déterminer les bornes inférieure et supérieure en nous appuyant sur la nature du verbe d'attribut EMBARK. En effet, ce verbe dénote une phase 2 de notre scénario, il est donc borné d'une part par la date la plus inférieure de celui-ci, et d'autre part par la date inférieure de tout événement appartenant à une phase postérieure - en l'occurrence LAND1. On remarquera que le verbe LAND, de l'événement LAND1 correspond à la phase 4 dans notre classification des étapes de voyage, la phase 3 n'étant pas décrite ici. Il signale la fin d'une première étape du voyage - et non pas la fin du scénario. Concernant les événements GO et LAND2, les bornes à attribuer ne peuvent être inférieures aux bornes de FLY. Comme cet événement est marqué avec les bornes supérieures du scénario, c'est finalement cette information qui est propagée à ces deux événements. On remarquera par ailleurs que si LAND2 traduit une nouvelle phase 4 de voyage du fait de la présence du verbe d'attribut LAND, l'événement GO, qui lui est concomitant correspond à la phase 5. De même que précédemment, le fait de nous limiter à une propagation locale de l'information limite la complexité à un $o(n^2)$.

5.2 Calcul des relations temporelles entre événements

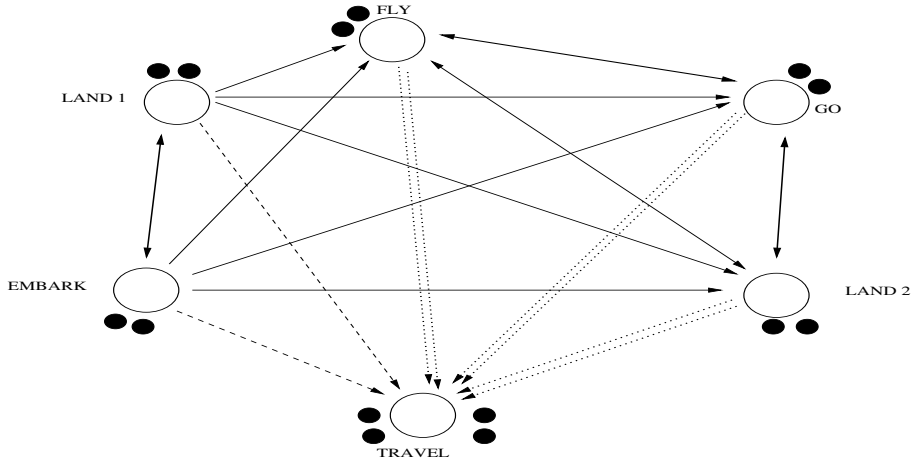

FIG. 6 – Graphe des événements complet

En utilisant le calcul des temporalités des événements, nous procédons à un tri sur les bornes des intervalles de temps de Allen, inspiré de l'algorithme proposé dans

[Mani *et al.*, 2003]. A partir des bornes triées, nous déterminons les relations temporelles entre tous les événements datés. La complexité du tri et du calcul des relations temporelles est un $o(n^3)$ où n est le nombre d'évènements de la dépêche. Pour simplifier cette tâche, nous ne tenons compte pour les intervalles d'intervalles que de la plus petite et de la plus grande des bornes. Ces résultats sont retranscrits en une liste pour chaque relation de Allen, liste qui énumère chaque couple d'événement partageant une telle relation. La figure 6 donne une représentation sous forme de graphe orienté des relations de Allen déterminée par le programme. Quatre types de relations d'Allen sont identifiés : EQUAL, BEFORE, FINISH et START. Les arêtes épaisses représentent la relation EQUAL, les arêtes simples la relation BEFORE, les arêtes en tirets la relation START, les arêtes doubles en pointillés la relation FINISH. Dans le cas d'un scénario de type TRAVEL, ce sont surtout les relations BEFORE qui sont pertinentes, la granularité et le calcul ne permettent pas d'approcher la précision de ces intervalles pour mettre en évidence des relations de type MEET. Les relations OVERLAP et DURING, ainsi que nous pouvons le constater, ne peuvent apparaître au sein d'un tel scénario, ce qui explique que ces balises soient vides. Les relations START et FINISH, pour ce type de scénario, ne peuvent être présentes que si un événement en encapsule d'autres, comme c'est le cas le cas pour TRAVEL. De tels événements encapsulés sont en relation soit de type START s'ils sont datés avec les bornes inférieures de TRAVEL (ce qui est le cas de EMBARK et LAND1), soit de type FINISH s'il correspondent à une borne supérieure de TRAVEL (ce qui est le cas ici avec FLY, LAND2 et GO). Concernant les relations de type EQUAL, on remarquera que les événements EMBARK et LAND1 d'une part et GO et LAND2 d'autre part ont été rangé abusivement dans cette catégorie. Cela est dû au fait que le programme qui calcule les relations de Allen a été implémenté avec une granularité qui n'est pas inférieure au jour. Ainsi, en étendant la représentation des intervalles d'Allen aux heures, minutes et secondes, il sera possible d'avoir une représentation qui positionne FLY et LAND2 dans une relation MEET, du fait que le verbe d'attribut FLY définit une phase 3 et que le verbe d'attribut LAND définit une phase 4 ce qui implique que ces événements sont consécutifs. De même LAND2 et GO sont consécutifs du fait de la balise TEMP-REL-WHEN marquant une succession immédiate des deux événements et devraient rentrer dans une relation MEET.

6 Expérimentation et résultats

6.1 Expérimentation

Nous avons étudié et annoté manuellement 36 dépêches issues du site PROMED-MAIL, corpus qui se décompose en 12 dépêches sur le virus Ebola, 12 dépêches sur l'épidémie de SARS, 8 dépêches sur Avian Influenza, et 4 dépêches sur des maladies non identifiées. Nous avons utilisé des transducteurs pour relever les termes lexicaux nouveaux dans 400 dépêches, dont 100 sur Ebola, 50 sur SARS, 70 sur Avian Influenza, et 180 concernant diverses maladies. Nous avons testé nos transducteurs sur 1230 dépêches, dont 337 sur Ebola, 182 sur SARS, 296 sur Avian Influenza, et 415 concernant diverses maladies. Nous avons procédé à un tirage au sort de 20 dépêches pour vérifier manuellement la pertinence de nos annotations et calculs. Nous procédons

à un traitement séquentiel des dépêches sous INTEX. A chaque étape du processus, le fichier produit est utilisé à l'étape suivante. Pour optimiser nos expérimentations, et réaliser ces tests dans le temps imparti, nous avons choisi manuellement à chaque étape les transducteurs à appliquer au texte pour annoter les séquences encapsulées dans des balises. Le protocole d'annotation (annoter une séquence la plus large possible, dont les composants sont identifiés par une inclusion de transducteurs, avant de traiter le contenu de cette séquence) nous garantit qu'une seule annotation est possible, et que les séquences à annoter seront traitées uniquement, et respectivement, par une variante du transducteur qui l'a reconnue initialement (cf. section 3.3). Après construction du document XML, nous le modifions avec notre application. Le traitement des dates incomplètes est réalisé en boucle. A chaque passage dans la boucle, une unique date incomplète est traitée, et le traitement s'arrête quand il ne reste plus de date incomplète ou qu'aucune modification n'a eu lieu lors d'un passage. Nous complétons alors les dates restantes avec les valeurs aux bornes. Nous trions enfin les dates des événements et déduisons de cet ordre les relations de Allen. Les résultats sont ajoutés en fin de document XML.

6.2 Résultats

Lors de l'expérimentation, nous n'avons pu appliquer les transducteurs dans leur totalité. En effet, le temps de calcul devenait rédhibitoire; de plus, certains de nos transducteurs dépassaient les limites de la machine et du logiciel. Il est très probable qu'avec la poursuite de l'extension des capacités des machines, de telles limites ne seront bientôt plus d'actualité. Notre approche sera donc exploitable pleinement. Les modifications apportées pour tenir compte des limites actuelles sont principalement dues à l'impossibilité de procéder à un étiquetage "en concurrence" avec tous les transducteurs. Nous avons donc modifié le processus de marquage afin d'appliquer manuellement et séquentiellement nos étiqueteurs les plus généraux. L'idéal serait d'avoir un transducteur "root" qui contienne les transducteurs pour chaque événement. Mais un tel transducteur, si on peut le dessiner, on ne peut ni le compiler ni l'appliquer à un texte pour l'instant. Dans l'immédiat nous envisageons donc de modifier l'application afin d'éviter une trop grande profondeur dans l'application des transducteurs. Cela se fera au prix, dans certains cas, d'une perte d'information dans l'annotation. Nous avons cherché à évaluer nos résultats en nous inspirant du protocole utilisé par [Mani et Wilson, 2000]. Rappelons que ces auteurs, à partir d'un ensemble de 221 articles obtiennent une annotation de 602 séquences temporelles correctement identifiées par rapport à un ensemble de 728 séquences obtenues humainement, ce qui donne une précision de 83,7 %. L'expérience réalisée sur 20 dépêches a donné plusieurs indications. Pour le marquage et l'annotation des événements, nous avons, après ajout des entrées lexicales inconnues, une annotation complète et sans erreur. Ce résultat de 100 % est à pondérer par le fait que seules certaines séquences spécifiques du texte sont annotées. De plus, les contraintes de temps ne nous ont pas permis d'analyser un échantillon suffisamment large de dépêches, ce qui nous incite à la prudence quant aux chiffres avancés à titre indicatif.

Le traitement des balises par notre application donne des résultats moins bons. Tous les intervalles sont complétés, mais le plus souvent avec des valeurs butoir (la majorité

des cas). De plus, moins d'un quart des intervalles complétés le sont correctement. Cela peut s'expliquer par le fait que nous ne relevons pas toutes les informations temporelles, mais seulement une partie de ces séquences, nous privant de points d'ancrage. Par ailleurs, de part le traitement réalisé, la probabilité pour un événement de se retrouver placé dans le fichier XML à proximité d'un événement de même type est augmentée. Dès lors, les scénarii prévus attribuent la même date à ces événements, et ce même si dans le texte d'origine ces deux événements étaient éloignés. Nous avons, à titre indicatif, ajouté manuellement des annotations pour les événements non marqués par notre traitement, et, dans ce cas, nous n'obtenons pas plus de résultats valides non triviaux, mais nous n'avons plus les erreurs précédentes. Dans une séquence comme "*He participated in a Congress in Hanoi between 11 and 17 May 2001. He fell ill during his travel back to New-York.*", nous sommes capables de déterminer que cette personne est tombée malade le 17 Mai, mais le rapprochement entre la date de fin de congrès et la date du retour nous est fourni par notre expérience : quand on se rend quelque part pour une réunion, habituellement, on repart de cet endroit après celle-ci. Ce genre de scénarii est loin d'être trivial à mettre en évidence et à caractériser. Les résultats des approches qui propagent la date localement sans distinction, notamment [Mani et Wilson, 2000], montrent que plus de la moitié des erreurs commises lors de tels traitements sont dues, justement, à cette propagation indifférenciée. Afin de se démarquer de ces travaux, nous avons choisi un traitement plus prudent, qui, au stade actuel de nos expériences, fournit des informations fiables. A titre indicatif, les résultats de l'étude citée montrent, à partir d'un petit échantillon de textes de 8505 mots, que 394 événements sur 663 sont propagés autour de verbes correctement étiquetés, ce qui donne 59,4% de données temporelles correctement annotées. Pour ce qui nous concerne, toujours à partir de notre échantillon de 20 notices, après propagation, sur un total de 217 séquences temporelles correctes, nous n'en produisons que 106, ce qui, dans des conditions similaires, nous donne une précision de 48,9% .

7 Conclusion

Le travail expérimental que nous avons réalisé présentait deux défis importants. Le premier était de donner une annotation des événements suffisamment riche et précise avec des grammaires locales implémentées sous forme de transducteurs. Le second était de montrer qu'à partir de ces données il était possible de faire des calculs temporels. C'est la raison pour laquelle nous avons utilisé le formalisme éprouvé des intervalles d'Allen. Afin de calculer et de propager des informations temporelles, complètes, incomplètes ou implicites, nous avons eu recours à des scénarios qui ont permis de donner une plus grande précision à l'information capturée. Nous considérons notre travail comme une étape préliminaire à l'utilisation de logiques temporelles. D'autre part, l'évaluation que nous avons menée montre que notre approche se situe, en termes de performances, au niveau d'autres travaux réalisés. Mais la principale limite de cette approche réside dans la capacité d'identifier convenablement et automatiquement les constituants d'une séquence d'événements temporellement liés [Johnston, 1994] [Moeschler, 1993]. En effet, nous nous sommes contentés dans ce travail, d'identifier des coréférences à une séquence d'événements de même type, situés à

proximité les uns des autres. Une nette amélioration de nos résultats pourraient être obtenue en réalisant de telles identifications sur une plus grande échelle, y compris entre différents textes. Nous pourrions par ailleurs élaborer des scénarii plus complexes, pour inférer de nouvelles relations temporelles entre événements de types différents. Nous disposerions ainsi de points d'ancrage supplémentaires, ce qui permettrait d'améliorer nos résultats. De même, l'étude de la périodicité des épidémies profiterait grandement d'une amélioration substantielle de la capacité à identifier des coréférences à un événement de même type.

Références

[Allen, 1983] J. F. Allen. Maintaining knowledge about temporal intervals. *Communication of the AC*, 26(11) :832–843, 1983.

[Allen, 1984] J. F. Allen. Towards a general theory of action and time. *Artificial Intelligence*, 23(2) :123–154, 1984.

[Borillo, 1998] A. Borillo. *L'espace et son expression en français.* Edition ORPHRYS, Paris, 1998.

[Bouaud, 2004] J. Bouaud. Abstraction de l'historique thérapeutique d'un patient par un traitement heuristique des indéterminations temporelles du dossier médical. In *In. RFIA '2004)*, Toulouse, 2004.

[Danlos, 1999] L. Danlos. Event Coreference Between Two Sentences. In *Proceedings of International Workshop on Computational Semantics, Tilburg, Pays-Bas*, 1999.

[Filatova, 2001] E. Filatova. Assigning time-stamps to event-clauses. In *Proceedings of the Workshop on Temporal and Spatial Reasoning at the Conference of the ACL*, Toulouse, 2001.

[Grishman *et al.*, 2002] R. Grishman, S. Huttunen, et R. Yangarber. Information extraction for enhanced access to disease outbreak reports. *Journal of Biomedical Informatics*, 35(4) :236–246, 2002.

[Johnston, 1994] M. Johnston. The role of aspect in the composition of temporal adverbial clauses with adverbs of quantification. In *Proceedings of the 25th Meeting of the North-Eastern Linguistics Society (NELS)*, GLSA, UMASS Amherst, 1994.

[Lascarides et Oberlander, 1993] A. Lascarides et J. Oberlander. Temporal connectives in a discourse context. In *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics*, pages 236–246, OTS - Research Institute for Language and Speech, 1993.

[Mani *et al.*, 2003] I. Mani, B.Schiffman, et J. Zhang. Inferring temporal ordering of events in news. In *Proceedings of the Human Language Technology Conference (HLT-NAACL'03)*, 2003.

[Mani et Wilson, 2000] I. Mani et G. Wilson. Robust temporal processing of news. In *Proceedings of the 38th Annual Meeting of the ACL*. Hong Kong, 2000.

[Moeschler, 1993] J. Moeschler. Aspects pragmatiques de la référence temporelle : indétermination, ordre temporel et inférence. *Langages*, 112 :39–54, 1993.

[Schilder, 1999] F. Schilder. Presupposition triggered by temporal connectives. In *Proceedings of the TALN '99, workshop Theoretical Bases for Semantics and Pragmatics in NLP*, pages 113–124, Cargese, France, 1999.

[Schilder, 2001] F. Schilder. From temporal expressions to temporal information : Semantic tagging of news messages. In *Proceedings of the ACL-2001, workshop on Temporal and Spatial Information Processing*, pages 65–72, Toulouse, 2001.

[Setzer et Gaizauskas, 2000] A. Setzer et R. Gaizauskas. Annotating events and temporal information in newswire texts. In *Proceedings of Second International Conference on Language Resources and Evaluation*, Athens, 2000.

Summary

We present in this paper a temporal tagging of events described in newswires from the PROMED MAIL epidemiologic website. After expounding the interest of tagging and temporal calculus from this textual data for epidemiologists, we present an experimentation using finite state transducers and Allen's Interval Relations. In order to tag events, we characterize part of the sublanguage of epidemiologic newswires and propose information extraction schemes implemented with transducers. Because it is not always possible to give a precise date related to an event, we propose a representation of event temporality based on the algebra of Allen temporal intervals as well as on its extensions. We use the relations described between these intervals to determine, from explicit temporal properties, the implicit temporal properties. We describe the implementation of scenarii enabling the extension of explicit temporal relations with a view to determining implicit temporalities. We propose a Java application that uses annotations obtained by transducers to conduct this process. Lastly, we present results obtained with corpora that are representative of a number of illnesses. The conclusion presents possibilities of expanding on our work.