Identification des composants temporels pour la représentation des dépêches épidémiologiques

Manal EL Zant[1], Liliane Pellegrin[1], Hervé Chaudet[1,2], Michel Roux[1]

[1]Laboratoire d'Informatique Fondamentale, UMR CNRS 6166
Équipe BIM, Faculté de médecine, 27 Bd Jean Moulin, 13005 Marseille
{el.zant, liliane.pellegrin, michel.roux}@medecine.univ-mrs.fr
[2]Unité de Recherche Epidémiologique, Département de Santé Publique,
IMTSSA, 13998 Marseille Armées
lhcp@acm.org
Date de la thèse (fin 2006)

Mots-clefs – Keywords

Analyse de textes, structure des évènements, extraction d'information, sémantique du temps

Text analysis, event structure, extraction information, temporal semantics

Résumé – Abstract

Dans le cadre du projet EpidémIA qui vise à la construction d'un système d'aide à la décision pour assister l'utilisateur dans son activité de gestion des risques sanitaires, un travail préalable sur la compositionalité des évènements (STEEL) nous a permis d'orienter notre travail dans le domaine de la localisation d'information spatio-temporelle. Nous avons construit des graphes de transducteurs pour identifier les informations temporelles sur un corpus de 100 dépêches de la langue anglaise de ProMed. Nous avons utilisé le système d'extraction d'information INTEX pour la construction de ces transducteurs. Les résultats obtenus présentent une efficacité de ces graphes pour l'identification des données temporelles.

EpidémIA project aims to the construction of a computerized decision-making system to assist user in his activity of medical risk management. A preliminary work on the events compositionality (STEEL) enables us to direct our work in the field of the space-time information localization. We have created some transducers graphs to identify temporal information on a corpus of 100 SARS ProMed English reports. We used the extraction information system INTEX to construct these transducers. The results obtained present an effectiveness of these graphs to identify temporal data.

1 Introduction

Les déplacements individuels dans un cadre professionnel ou de loisir et l'essor des nouvelles épidémies sur la planète ont conduit à la création d'un nombre croissant de dispositifs de veille sanitaire dont la liste de diffusion internationale PROMED (http://www.promedmail.org). Simultanément du fait d'Internet, nous assistons à ladématérialisation généralisée de l'information utilisable pour la veille épidémiologique. Dans ce cadre, le projet EpidémIA a pour objectif le traitement des dépêches de ProMed pour l'analyse, la modélisation, la gestion et la restitution de connaissances et des données épidémiologiques. Un travail préalable nous a permis de développer un langage formel de représentation des connaissances (STEEL) adapté à la problématique des informations épidémiologiques, tenant compte à la fois de l'orientation évènementielle des récits, de la compositionalité des évènements et de leur localisation spatio-temporelle (Chaudet, 2004). Le travail que nous présentons ici concerne l'aspect TALN de projet, et en particulier la mise au point d'une méthode automatique d'extraction d'information de ces composants afin de pouvoir les utiliser pour l'inclure dans le modèle STEEL. Dans ce travail, nous abordons en particulier le problème d'identification et d'extraction des expressions temporelles en créant des automates INTEX (Silberztein, 1993). Orienté par STEEL, nos graphes de transducteurs recherchent les séquences possibles, dans le corpus des dépêches, contenants les éléments sémantiques nécessaires à la construction d'une représentation temporelle des évènements.

Dans cet article nous présentons d'abord le modèle STEEL sur lequel reposera la configuration des transducteurs INTEX. Nous dressons ensuite un état de l'art sur les différentes approches d'annotation temporelles en les comparant à la nôtre. Une étude des expressions temporelles est ensuite effectuée sur notre corpus, formé de 100 dépêches (248670 mots) de l'épidémie de SRAS (partie 4). Ceci nous permettra de proposer un graphe temporel spécifique tout en présentant et commentant les résultats obtenus.

2 Le modèle STEEL

Dans le cadre de la modélisation des systèmes dynamiques, plusieurs formalismes adaptés au raisonnement actions/évènements et leurs effets ont été proposés. Le calcul d'évènement de Kowalski et Sergot (1986) est un de ceux-ci. En se fondant sur cette approche et celle de Cervesato et Montanari (2000), Chaudet (2004) a créé une adaptation pour la représentation des récits épidémiologiques, STEEL (Spatio-Temporal Extended Event Language), qui se caractérise par la possibilité de représenter des agrégats d'évènements spatio-temporellement localisés. Les entités temporelles et spatiales y sont réifiées et introduites comme arguments de prédicats spécifiques dans une théorie de premier ordre. De plus le langage STEEL intègre les primitives de manipulation des évènements. Pour simplifier, au lieu de la forme traditionnelle *Happens(action,temps)*, les évènements selon STEEL se produisent dans un lieu spécifique : *Happens(macro-évènements,temps,espace)*. Trois composantes du discours doivent donc être identifiées et représentées de façon coordonnée dans le langage de représentation : l'évènement (simple ou complexe), le temps et le lieu. Par exemple, pour la phrase *The total number of dengue-affected patients, according to the official account, stood at 4763, as of 16 Sep 2003. Of these, 45 have died so far,* STEEL donne:

happens (e1,<t1, Bangladesh>, <2002-09-16, Bangladesh>)
happens (e2, <t2, Bangladesh>, <2002-09-16, Bangladesh>)

instance (e1,macroevent); instance (e2, macroevent)
meventdef(e1, iterevent(infection,4763)) ; meventdef(e2, iterevent(death,45))
agent (e1, dengue)
experiencer (e2,a) → experiencer (e1,a)
$t1 \leq 2002\text{-}09\text{-}16 \wedge t1 \leq t2 \wedge t2 \leq 2002\text{-}09\text{-}16$

3 Localisation temporelle

L'annotation précise et détaillée des expressions temporelles a commencé avec les conférences MUC 5-7 (Message Understanding Conferences) pour l'identification et la classification des entités nommées EN (Chinchor, 1999). Dans la même vision, Ferro et al. (2001) décrivent un ensemble de consignes pour l'annotation des expressions temporelles à partir de plusieurs langues, et leur associent une représentation canonique du temps à laquelle elles se réfèrent. Cependant, une autre approche de l'annotation a aussi été utilisée. C'est le marquage temporel, qui vise à associer un temps du calendrier à certains ou tous les événements du texte. Filatova et Hovy (2001) décrivent une méthode pour fractionner des phrases en leurs événements constitutifs et leur assigner des marqueurs temporels. Le marquage utilise deux temps principaux : le temps de l'article et le dernier temps indiqué dans la même phrase. Dans cette approche Schilder et Habel (2001) ont développé un système d'étiquetage sémantique des expressions temporelles. Elles sont classifiées selon deux types : celles qui se rapportent à un temps du calendrier ou d'horloge et celles qui se rapportent à des événements. L'ensemble des relations temporelles proposées est équivalent aux relations d'Allen (1983). Une troisième approche (Setzer et Gaizauskas, 2000) se focalise sur les relations temporelles entre les événements et le temps ou entre les événements mêmes. Cette approche prend l'identification des relations temporelles comme but et repose sur la façon dont l'information temporelle se présente ainsi que sa relation avec le texte. Leur schéma permet de déterminer l'ordre relatif ou le temps absolu des évènements. Katz et Arosio (2001) à leur tour ont proposé une annotation des informations des relations temporelles en se basant sur les relations entre évènement. Notre approche se rattache à la deuxième catégorie de travaux. L'annotation est pratiquée en tenant compte de la date de la dépêche et de celle signalée dans le récit. L'association entre l'événement et le marquage temporel se fait ultérieurement au niveau de la représentation logique.

4 Méthodologie

Notre approche permet d'analyser des membres de phrases qui peuvent comporter une expression temporelle. Dans ce cadre, nous l'avons décomposée en 6 étapes :

1 Isoler les mots caractérisant la localisation temporelle des évènements épidémiques,
2 Construire des dictionnaires[1] spécifiques pour ces mots. Par exemple, pour *April* la forme sera : *Apr,N+Month* et pour *Hong Kong* : *Hong Kong,N+Location*,etc.
3 Sélectionner dans le corpus les membres de phrases comportant un élément temporel par une recherche automatique (INTEX) des mots typés à l'étape 2,

[1] Un dictionnaire INTEX est une liste de termes avec une étiquette et une liste d'information sémantique associée aux informations flexionnelles et lexicales

4 Analyser la configuration syntaxique et sémantique qui entoure l'élément temporel. INTEX construit les graphes syntaxiques correspondant aux membres de phrases sélectionnés. Mais, il donne plusieurs solutions. Les ambiguïtés sont nombreuses, dues aux dictionnaires non spécifiques du domaine. Il est donc nécessaire de construire des dictionnaires spécifiques,

5 Elaborer les graphes correspondants. Ces graphes sont utilisés pour les formes complexes où des phénomènes d'insertion et d'optionalité interviennent. Nous donnons en figure 1 un exemple de transducteur[2] reconnaissant les numéro du jour ainsi que le jour en caractère. Dans ce schéma, la première partie <N+DateJour> désigne dans les dictionnaires de temps les différentes formes des noms des jours de la semaine comme (Monday, Mon, mon, MONDAY, etc.). La deuxième partie Day_num est le sous graphe qui est constitué des numéros des jours d'un mois (figure 2),

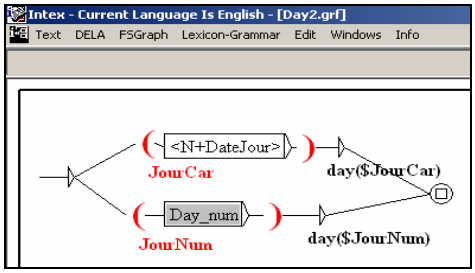

Figure1-Transducteur de Jour

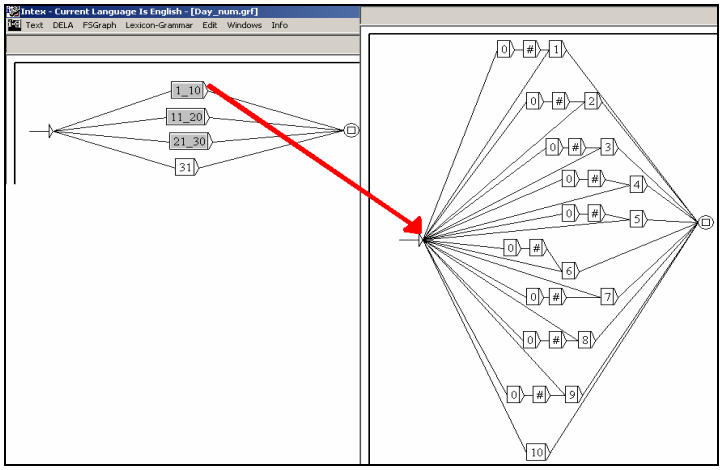

Figure 2-Numéro des jours du mois

6 Tester leurs performances en évaluant le taux d'erreur : rapport entre le nombre de séquences reconnues erronées sur le nombre total des séquences identifiées.

5 Résultats et discussion

Afin de créer les graphes de localisations temporelles, nous avons étudié les différentes formes présentes dans notre corpus formé de 100 dépêches de l'épidémie du SRAS. Elles ont été regroupées selon deux catégories principales.

Un graphe de transducteurs a été construit pour identifier ces expressions temporelles (Figure 3). En particulier, l'étiquette **"Month($MoisNum)"** provoque l'écriture du prédicat **"Month"** avec un argument: la valeur de la variable **"MoisNum"**, valeur trouvée dans le texte. Dans l'échantillon de 100 dépêches, 6284 séquences sont reconnues par ce graphe. Le tableau 1 présente quelques séquences reconnues, ainsi que leurs équivalents en mode de remplacement. Premièrement, des formes langagières spécifiques des dépêches ont été identifiées. Il s'agit de plusieurs formats non littéraires, de style rédactionnel abrégé comme les expressions suivantes, *10 Apr 2003, April 10, 2003, 10 April 2003, 2003_03_10, 2003-*

[2] Un transducteur est un automate qui reconnaît des séquences de mots et peut produire une nouvelle séquence

04/10, Friday [10 Apr 200, 10 Apr 2003, 2003/04/10, 10-Apr-2003, 10 April 2003, April 10, 2003, etc. Deuxièmement, les cas d'expressions temporelles présentant des formes littéraire plus classiques ont été identifiées dans notre corpus comme *In mid February, after several days, in recent days, during the second week of February, Monday evening, last Friday night.*

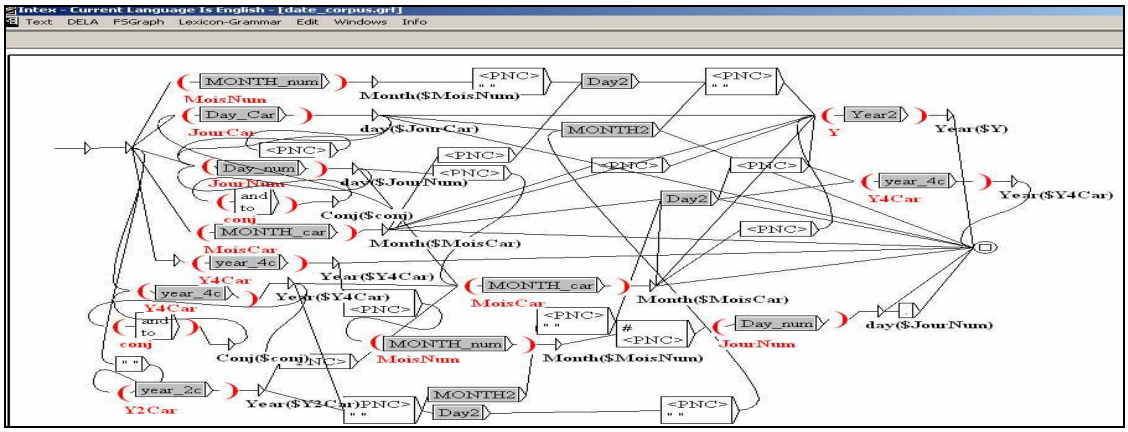

Figure 3- Annotation des expressions temporelles

Séquences reconnues	***Résultat en Mode de remplacement***
Worldwide 20030315.0637 …	Worldwide Year(2003)Month(03)day(15) 0637
Friday [18 Apr 2003]	day(Friday (day(18)Month(Apr)Year(2003)]
Monday [21 Apr 2003]	day(Monday)day(21)Month(Apr)Year(2003)]

Tableau 1. Exemples de séquences reconnues

Pour cela, nous avons bénéficié de la bibliothèque de graphes de Maurice Gross disponible sur le site d'INTEX. Appliqués sur ces dépêches, ces graphes identifient les formes littéraires des expressions temporelles dans ce corpus. Cependant, ils identifient à tort d'autres expressions comme *that **may** have identified, **from 16** countries, **in a second** Hong Kong hospital, **in terms** of industries, **on the 9th** floor.* Pour réduire ces cas d'erreurs, nous avons modifié les graphes concernés. Ils s'adaptent mieux au langage professionnel utilisé dans les dépêches.

Le graphe qui englobe finalement l'ensemble des formes associées aux expressions temporelles est composé des deux graphes décrits précédemment. 7087 cas ont pu être identifié dans notre corpus par l'application de ce graphe. Il reconnaît des expressions littéraires et non littéraires. Nous présentons dans le tableau 2 quelques séquences reconnues, ainsi que leurs résultats en mode de remplacement.

Nous avons décrit dans cet article une expérimentation d'utilisation d'INTEX afin d'extraire les expressions temporelles. Elle a permis d'extraire toutes les expressions temporelles, avec un taux d'erreur de 3 sur 6284 formes. Une solution serait, afin d'améliorer l'identification de ces différentes expressions, de passer à une forme générale, qui serait obtenue grâce à la fonction de génération des transducteurs. On substituerait à toutes séquences d'origine, des séquences générées à partir de marqueurs définis (figés) dans les graphes et de mots ou

marqueurs sémantiques récupérés par INTEX dans des variables, à partir du texte d'origine. Dans une étape ultérieure, cette forme devrait être traduite dans le langage STEEL.

Séquence Reconnues	***Résultat en mode de remplacement***
in the evening on 10 Apr 2003	ExpTemp(in the evening)day(10)Month(Apr))Year(2003)
the previous month [February]	ExpTemp(the previous month)Month(February)
yesterday (28 Apr 2003)	ExpTemp(yesterday)day(28)Month(Apr)Year(2003)

Tableau 2. Exemples de séquences reconnues

Références

Allen J.F. (1983), Maintaining Knowledge about Temporal Intervals, *Communications of the ACM,* vol 26(11), pp. 832-843.

Cervesato I., Montanari A. (2000), A Calculus of Macro-Events: Progress Report. *In 7th International Workshop on Temporal Representation and Reasoning*, 47-58.

Chaudet H. (2004), Une extension du Calcul des Evènements pour la représentation de récits épidémiologiques, *15ème journée francophone IC'2004*, 285-296.

Chinchor N., Brown E., Ferro L., Robinson P. (1999), Named Entity Recognition Task Definition, *MITRE.*

Ferro L., Mani I., Sundheim B., Wilson G. (2001), TIDES Temporal Annotation Guidelines Draft - Version 1.02, *MITRE Technical Report MTR 01W000004.* McLean, Virginia.

Filatova E., Hovy E. (2001), Assigning Time-Stamps to Event-Clauses, *In Proceedings of the ACL-2001 Workshop on Temporal and Spatial Information Processing.*

Katz G., Arosio F. (2001), The Annotation of Temporal Information in Natural Language Sentences. *In Proceedings of the ACL-2001 Workshop on Temporal and Spatial Information Processing*, 104-111.

Kowalski R., Sergot M. (1986), A Logic-based Calculus of Event, *New Generation Computing*, Vol.4 , pp.67-95.

Schilder F., Habel C. (2001), From Temporal Expressions To Temporal Information: Semantic Tagging Of News Messages, *In Proceedings of the ACL-2001 Workshop on Temporal and Spatial Information Processing,* 65-72.

Setzer A., Gaizauskas R. (2000), Annotating Events and Temporal Information in Newswire Texts. *In Proceedings of the Second International Conference on Language Resources and Evaluation*, 1287-1294.

Silberztein M. (1993), *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX.* Masson: Paris.