Une méthode d'extraction d'information fondée sur les graphes pour le remplissage de formulaires

Ludovic Jean-Louis — Romaric Besançon — Olivier Ferret

CEA, LIST, Laboratoire Vision et Ingénierie des Contenus
91191 Gif-sur-Yvette Cedex, France
{ludovic.jean-louis,romaric.besancon,olivier.ferret}@cea.fr

RÉSUMÉ. *Une tâche importante des systèmes d'extraction d'information se focalisant sur des événements est le remplissage de formulaires regroupant, en les caractérisant par leur type, les informations associées à un événement donné à partir d'un texte. Cette tâche peut s'avérer difficile lorsque l'information est dispersée à l'échelle du texte et mélangée à des éléments d'information liés à d'autres événements similaires. Nous proposons dans cet article une approche en deux étapes pour prendre en compte ce problème : d'abord une segmentation du texte en événements pour identifier les phrases relatives à un même événement, puis une méthode de sélection des entités liées à l'événement dans ces phrases. Une évaluation de cette approche sur un corpus annoté de dépêches dans le domaine des événements sismiques montre une F1-mesure de 77 % pour la tâche de remplissage de formulaires.*

ABSTRACT. *In event-based Information Extraction systems, a major task is the filling from a text of a template gathering information related to a particular event. Such template filling may be a hard task when the information is scattered throughout the text and mixed with similar pieces of information relative to different events. We propose in this paper a two-step approach for template filling: first, an event-based segmentation is performed to select the parts of the text related to the target event; then, a graph-based method is applied to choose the most relevant entities in these parts for characterizing the event. An evaluation of this model based on an annotated corpus for earthquake events shows a 77% F1-measure for the template-filling task.*

MOTS-CLÉS : *extraction d'information, segmentation événementielle, remplissage de formulaires.*

KEYWORDS: *information extraction, event-based text segmentation, template filling.*

TAL. Volume 54 – n° 1/2013, pages 139 à 170

1. Introduction

Le domaine de l'extraction d'information couvre toutes les tâches consistant à extraire des informations structurées à partir de textes. Une tâche archétypique de ce domaine est celle définie dans les conférences MUC (*Message Understanding Conferences*) (Grishman et Sundheim, 1996), où les systèmes doivent permettre de remplir de façon automatique des formulaires (ou *templates*) concernant des événements. Il n'est pas facile de donner une définition stricte de la notion d'événement mais, à la suite de Arnulphy (2012), on peut considérer qu'il s'agit d'une notion caractérisant que quelque chose change, associant en particulier « mouvement, insertion dans le temps, modification d'une situation ». Du point de vue de l'extraction d'information, un événement est aussi caractérisé par les informations spécifiques qui lui sont associées. Les formulaires permettent de mettre en évidence les informations spécifiques à un type d'événement considéré et d'ignorer tout autre type d'information non pertinente. La figure 1 présente un exemple du remplissage d'un formulaire relatif aux tremblements de terre à partir du texte d'une dépêche de presse.

Texte	***Formulaires***
EV1Un **séisme** de magnitude **7,2** sur l'échelle de Richter a frappé **samedi** la ville de **Kurihara** (**préfecture de Miyagi**). EV1Le **tremblement** s'est produit à **08H43**, heure locale. EV1La **secousse** a été ressentie jusqu'à **Tokyo**, à 500 kilomètres au sud des préfectures japonaises d'**Iwate** et de **Miyagi**, principales zones touchées. Les **séismes** sont courants au **Japon**, qui est l'une des zones sismiques les plus actives de la planète. EV2En **octobre 2004**, un **séisme** d'une magnitude de **6,8** avait touché la région de **Niigata**, dans le nord du pays.	*EV1* • **ÉVÉNEMENT** : séisme, tremblement, secousse • **DATE** : samedi • **HEURE** : 08h43 • **MAGNITUDE** : 7,2 • **LIEU** : Kurihara *EV2* • **ÉVÉNEMENT** : séisme • **DATE** : octobre 2004 • **HEURE** : / • **MAGNITUDE** : 6,8 • **LIEU** : Niigata

Figure 1. *Exemple de remplissage de formulaires*

La réalisation d'un système complet d'extraction d'information pour le remplissage de formulaires présente plusieurs difficultés, parmi lesquelles l'identification des entités nommées ou des autres entités spécifiques du domaine, l'établissement des relations entre ces entités, la résolution de la coréférence concernant les entités, le regroupement d'informations dispersées dans le texte, etc. (Turmo *et al.*, 2006).

Il n'existe pas actuellement d'approche considérée comme standard pour le remplissage de formulaires. Néanmoins, la plupart des systèmes d'extraction d'information adoptent une approche en deux temps : des patrons spécifiques au domaine ou des classifieurs sont d'abord utilisés pour extraire au niveau phrastique les informations constitutives du formulaire considéré (dates, lieux, magnitudes et heures dans le

cas des événements sismiques de la figure 1) en s'appuyant sur les mentions d'événements ; des heuristiques relatives au type d'événement ou au type de textes considéré sont ensuite appliquées pour fusionner les informations extraites dans des formulaires globaux. Même si ce type d'approche est largement utilisé, il se heurte à deux problèmes importants : une vision très locale de l'extraction des informations élémentaires et une prise en compte limitée et peu générique des dépendances entre ces informations, en particulier pour le remplissage des formulaires.

La figure 1 illustre clairement le fait que les informations relatives à un événement, ici *EV1*, peuvent être exprimées au-delà de la portée de la phrase. Ce problème pose plus généralement la question de la délimitation des parties des textes relatives à un événement ou un type d'événement donné car les informations d'un événement ne sont pas toujours liées à une mention d'événement proche. Notre approche pour résoudre ce problème s'appuie sur une segmentation discursive des textes sur la base des événements auxquels chaque phrase fait référence. Plus largement, son objectif est de diminuer l'espace textuel à explorer pour faire le lien entre une entité et une mention d'événement et donc *in fine*, pour le remplissage du formulaire associé à un événement donné. Par définition, un événement est ancré dans le temps, et des cadres temporels distincts ne peuvent pas correspondre à un même événement. Nous choisissons donc de fonder cette segmentation événementielle sur des indices de nature temporelle.

Le second problème évoqué ci-dessus a déjà fait l'objet de quelques travaux assimilant les formulaires à des relations complexes. Dans ce contexte, chaque événement est vu comme une relation n-aire dont l'arité est égale au nombre de champs à remplir dans le formulaire ($n = 5$ dans l'exemple précédent). Cette vision a d'abord été appliquée au niveau local pour des phrases contenant plusieurs entités d'intérêt pour le même événement (la première phrase de la figure 1 en contient par exemple quatre) : dans (McDonald *et al.*, 2005), les relations entre la mention de cet événement et les informations qui lui sont liées ne sont ainsi plus considérées indépendamment les unes des autres mais de façon plus globale. Au-delà, plusieurs méthodes ont été proposées pour extraire des relations complexes, parmi lesquelles se distinguent des méthodes à base de graphe (McDonald *et al.*, 2005 ; Wick *et al.*, 2006) et des méthodes à base d'inférences (Goertzel *et al.*, 2006). Dans cet article, nous présentons une méthode à base de graphe, en commençant par construire un graphe d'entités fondé sur le résultat de la segmentation et en utilisant plusieurs stratégies génériques (*i.e.* indépendantes du domaine considéré) pour la construction de la relation complexe à partir de ce graphe.

La suite de l'article est composée comme suit : la section suivante présente un état de l'art sur le sujet, la section 3 présente une vue générale de notre approche. Nous décrivons ensuite plus en détail les méthodes utilisées pour les étapes de segmentation et de remplissage de formulaires dans les sections 4 et 5. Finalement, une évaluation de l'approche est présentée à la section 6 pour les deux étapes, suivie d'une discussion à la section 7, qui détaille également de façon plus précise les différences de notre approche par rapport à des travaux comparables.

2. Motivation et état de l'art

Le remplissage de formulaires est une tâche centrale des systèmes d'extraction d'information et a fait l'objet en tant que telle de nombreuses études. Ainsi, dans le contexte des campagnes d'évaluation MUC (*Message Understanding Conferences*) et ACE (*Automatic Content Extraction*) (Doddington *et al.*, 2004), un des objectifs des systèmes participants était de remplir automatiquement des formulaires prédéfinis avec une structure fixe. Bien que ce soit l'approche la plus répandue, d'autres travaux, comme ceux de (Chambers et Jurafsky, 2011), adoptent un point de vue différent et proposent une approche non supervisée pour remplir des formulaires sans connaissance *a priori* sur leur structure. Ils exploitent dans ce cas des techniques de regroupement (*clustering*) pour apprendre la structure des formulaires et des patrons syntaxiques pour en remplir les champs.

Une grande partie des systèmes d'extraction d'information, en particulier ceux fondés sur des approches à base d'apprentissage automatique, s'appuient sur l'idée qu'un événement est souvent décrit dans une seule phrase, ce qui conduit à donner une importance moindre à l'information interphrastique. Cette idée est nommée « hypothèse de la phrase seule » (*single sentence assumption*) par Stevenson (2006), qui rapporte qu'au plus 60 % des faits mentionnés dans les corpus MUC (MUC 4, 6 et 7) peuvent être identifiés avec cette hypothèse. Ce pourcentage a été confirmé plus récemment par Ji *et al.* (2010), montrant qu'environ 40 % des relations entre entités dans le corpus de la campagne TAC-KBP nécessitent l'usage de techniques d'inférences interphrastiques pour les extraire.

Peu d'approches ont été proposées pour faire de l'extraction d'information à un niveau discursif sans reposer sur des heuristiques liées au domaine abordé. Parmi elles, celles de Gu et Cercone (2006) et Patwardhan et Riloff (2007) sont les plus proches de l'approche présentée ici. Gu et Cercone (2006) définissent une approche à base de modèles de Markov cachés, d'une part pour identifier les unités de textes (phrases) pertinentes pour le remplissage de formulaires, et d'autre part pour faire l'extraction des entités dans les phrases retenues. De façon similaire, Patwardhan et Riloff (2007) proposent tout d'abord d'identifier les phrases pertinentes en utilisant un modèle SVM (*Support Vector Machine*), puis d'appliquer différents niveaux de patrons d'extraction pour remplir les champs du formulaire.

Une partie des travaux sur le remplissage des formulaires considère ces derniers comme des relations *n*-aires. Une des premières approches ayant adopté ce point de vue vient du domaine biomédical (McDonald *et al.*, 2005) mais se limitait au cadre phrastique. Elle a ensuite été appliquée dans le domaine des mouvements de personnel dans les entreprises (Afzal, 2009), avec un cadre plus large. D'autres travaux se sont attaqués au problème des relations complexes dans le contexte de l'extraction de champs pour les bases de données (*database record extraction*), en s'intéressant plus particulièrement à la compatibilité d'un ensemble d'entités données plutôt que d'une paire d'entités, ce qui les a amenés à prendre en compte des relations interphrastiques entre entités (Wick *et al.*, 2006 ; Mansuri et Sarawagi, 2006 ; Feng *et al.*, 2007).

Ce rapide aperçu des travaux existants concernant le remplissage de formulaires met en évidence un double constat. En premier lieu, ces travaux font pour une large part l'hypothèse que les informations concernant un événement donné apparaissent dans l'environnement immédiat de la mention de cet événement. Cette hypothèse de localité permet de faire abstraction des problèmes d'ambiguïté de rattachement d'une entité lorsque plusieurs événements de même type sont évoqués dans un texte. Quand ils ne font pas une telle hypothèse, ces travaux s'appuient fréquemment sur des moyens d'identification des informations à extraire dépendant de façon plus ou moins étroite du domaine considéré. Au travers du travail que nous présentons dans cet article, nous tentons de dépasser cette double limite en prenant en compte explicitement le fait que les informations relatives à un événement peuvent être dispersées à l'échelle d'un texte et que ce même texte peut évoquer plusieurs événements du même type. Cette prise en compte est en outre réalisée en s'appuyant autant que possible sur des processus d'analyse générique, sans s'affranchir complètement d'une dépendance au domaine au travers de l'exploitation de corpus annotés.

3. Description de l'approche et de son contexte

L'approche d'extraction d'événements présentée dans cet article prend place dans un cadre applicatif de veille où les utilisateurs ne sont en général intéressés que par les événements les plus récents. Plus précisément, les événements considérés dans ce cadre sont des événements à caractère sismique, décrits par le formulaire du tableau 1. Ce formulaire spécifie les différents types d'information recherchés pour chaque événement sismique, le type TYPE_EVT ayant un caractère un peu particulier puisque son identification est étroitement impliquée avec la détection des événements (il correspond de fait à la mention de l'événement dans le texte).

Rôle	Nature
TYPE_EVT	type d'événement (séisme, tsunami ...)
LIEU	lieu de l'événement
DATE	date de l'événement
HEURE	heure de l'événement
MAGNITUDE	magnitude
COORD_GEO	coordonnées géographiques de l'événement (longitude/latitude)
DOMMAGES	dégâts causés par l'événement

Tableau 1. *Formulaire de description des événements sismiques sous-tendant l'exemple de la figure1*

Dans ce contexte, notre but est de synthétiser, à partir de dépêches de presse collectées à partir du Web, les informations relatives aux événements récents dans un tableau de bord. Néanmoins, les dépêches considérées font souvent référence à plusieurs événements comparables, à l'instar de l'exemple de la figure 1, généralement pour mettre en évidence les similarités ou les différences entre l'événement récent et

des événements passés de même nature. Dans notre application spécifique de veille, nous ne nous intéressons pas aux événements passés, que nous considérons comme une source de bruit pour la détection des informations relatives à l'événement principal de l'article. Nous avons donc fait l'hypothèse, comme Feng *et al.* (2007), qu'un document est associé à un seul formulaire, celui-ci décrivant l'événement principal du document. Nous avons par ailleurs défini une stratégie en deux étapes pour extraire cette information en tenant compte de la complexité de la structure des textes considérés (Jean-Louis *et al.*, 2011a) :

– une segmentation des textes en événements : les informations relatives à l'événement principal d'un texte peuvent être réparties sur plusieurs phrases. Par conséquent, nous avons choisi de découper chaque texte en segments homogènes du point de vue événementiel afin de focaliser l'espace de recherche des informations caractérisant l'événement principal. Ces segments regroupent fréquemment des phrases non contiguës car la structure des articles fait souvent des allers-retours entre l'événement principal et un ou plusieurs événements passés ;

– le remplissage des formulaires : cette étape est réalisée au sein des segments événementiels liés à l'événement principal et a pour objectif de sélectionner les entités de ces segments véritablement liées à l'événement principal parmi les entités candidates, sélectionnées par leur type. Cette sélection s'appuie conjointement sur les relations entre entités identifiables au niveau phrastique et sur les relations intervenant à l'échelle textuelle.

Un formulaire synthétisant les informations relatives à l'événement principal de chaque dépêche est ainsi construit et intégré dans le tableau de bord. Par ailleurs, cette approche d'extraction d'information s'inscrit dans un cadre plus général d'une application, qui compte en particulier en amont un système de filtrage permettant de sélectionner les documents ayant pour objet principal un événement à caractère sismique. Ce filtrage est classiquement mis en œuvre dans notre cas par un classifieur statistique selon les modalités décrites dans (Besançon *et al.*, 2012). Même si les articles sont initialement sélectionnés par le biais de requêtes constituées de mots-clés en lien avec les événements sismiques, l'absence de filtrage conduirait à une situation dans laquelle 60 % des articles ne seraient pas pertinents, avec pour conséquence une diminution de moitié des performances du processus d'extraction d'information. Les expérimentations présentées à la section 6 se focalisant plus spécifiquement sur ce processus d'extraction d'information, tous les documents constituant le corpus d'évaluation ont été sélectionnés manuellement et sont donc pertinents. D'autre part, la stratégie exposée reste cantonnée à un cadre monodocument : les informations extraites de différents textes mais caractérisant un même événement ne sont ni recoupées, ni fusionnées. Néanmoins, de premières expériences sur le regroupement *a posteriori* des dépêches ont montré que l'exploitation des informations extraites dans les formulaires permet d'améliorer ce regroupement (Besançon *et al.*, 2012).

4. Segmentation événementielle des textes

L'idée de segmenter des textes en unités homogènes du point de vue événementiel a principalement été abordée selon deux angles : soit avec des méthodes reposant sur des modèles très lexicalisés, de façon assez liée à un domaine particulier (Gu et Cercone, 2006 ; Patwardhan et Riloff, 2007), soit en ne s'appuyant que sur la logique d'enchaînement des types d'événements (Naughton, 2007). Notre approche est intermédiaire : en exploitant des informations de nature temporelle, elle fait appel à des caractéristiques des textes dépassant leur simple appartenance à un domaine donné (Jean-Louis *et al.*, 2010).

Du point de vue du processus de segmentation, un texte est vu comme une séquence de phrases, chaque phrase étant caractérisée par un statut événementiel. Comme dans la plupart des travaux similaires, nous faisons l'hypothèse, en pratique raisonnablement simplificatrice, qu'une phrase possède un statut événementiel homogène. Nous distinguons plus précisément trois statuts :

– **événement principal** : la phrase fait référence à l'événement principal du texte ;

– **événement secondaire** : la phrase fait référence à un événement secondaire du texte, sans distinction de l'événement particulier s'il en existe plusieurs ;

– **contexte** : la phrase ne fait référence à aucun événement du type considéré.

Dans cette perspective, nous considérons la segmentation événementielle comme une tâche de classification visant à associer à chaque phrase d'un texte un statut événementiel. Néanmoins, une telle segmentation possède un caractère intrinsèquement discursif dans la mesure où les catégories événementielles ne s'enchaînent pas de manière arbitraire. Du point de vue de la classification des phrases, elles sont donc déterminées à la fois par des indices repérables au niveau phrastique mais également par les catégories et les indices des phrases précédentes. Compte tenu de l'étroite dépendance existant entre les dimensions temporelle et événementielle des textes (Pustejovsky *et al.*, 2005), nous avons choisi de nous appuyer sur des indices de nature temporelle. Notre approche se focalise donc sur la capture des relations entre les changements événementiels et les changements de cadre temporel, se manifestant par exemple au travers du passage du passé composé vers le plus-que-parfait accompagnant la transition de l'événement principal à un événement secondaire dans le texte de la figure 2. Cette figure met également en avant d'autres indices temporels intéressants de ce point de vue, comme la présence des dates. Le choix de s'appuyer sur la dimension temporelle des textes pour mettre en évidence leur structure événementielle va par ailleurs dans le même sens que des travaux tels que (Eberle, 1992), même si dans notre cas nous ne nous inscrivons pas dans un cadre formel comme celui de la *Discourse Representation Theory* (DRT).

Sur le plan technique, nous avons choisi de traiter la problématique de la segmentation événementielle des textes comme un problème de classification de séquences. Nous avons de ce fait testé deux modèles de référence pour ce type de tâches : les mo-

{PRINC} Un tremblement de terre de 5,6 sur l'échelle de Richter a frappé Jayapura, Papua, **dimanche** peu de temps après minuit.
{SEC} **Samedi**, l'agence avait précédemment enregistré un tremblement de terre de magnitude 5,6 qui avait frappé Melonguane, dans le nord du Sulawesi.
{CONT} L'indonésie est située au-dessus d'une zone vulnérable aux tremblements de terre appelée la ceinture de feu du Pacifique.

Figure 2. *Exemple de segmentation d'un texte : {*PRINC*} = événement principal, {*SEC*} = événement secondaire, {*CONT*} = contexte*

dèles de Markov cachés (*Hidden Markov Model*, ou HMM) et les champs aléatoires conditionnels (*Conditional Random Fields*, ou CRF).

4.1. *Prétraitement des dépêches*

Une étape préliminaire à la segmentation des textes en événements est le repérage des informations temporelles à la base de notre approche. Pour ce repérage nous appliquons à chaque texte la chaîne de traitements linguistiques suivante : tokenisation, détection des fins de phrases, désambiguïsation morphosyntaxique, identification du temps des verbes, reconnaissance d'expressions figées ou semi-figées[1] et reconnaissance des entités nommées. Cette chaîne de traitements est mise en œuvre par l'analyseur linguistique LIMA présenté dans (Besançon *et al.*, 2010).

4.2. *Modèle HMM*

Les HMM constituent un modèle de classification de séquences (Rabiner, 1989) très largement utilisé en traitement automatique des langues (TAL) dans des tâches telles que la reconnaissance d'entités nommées ou la désambiguïsation morphosyntaxique, mais également la segmentation de textes, comme par exemple dans le cas de la segmentation thématique (Yamron *et al.*, 1998). Les HMM sont des automates stochastiques à états finis permettant de déduire des séquences d'états non observables (ou états cachés) à partir de séquences de données observées (observables). Dans notre cas, nous cherchons à déterminer la séquence d'événements associée à la séquence de phrases formant un texte donné.

Nous faisons l'hypothèse que cette segmentation est un processus markovien, c'est-à-dire que l'état associé à l'observable courant ne dépend que de celui-ci et de l'état précédent : nous utilisons les marqueurs temporels, dans le cas présent les temps grammaticaux, comme observables, les catégories d'événements constituant les états cachés. Les matrices de transition (une pour les états, une pour les observables) sont

1. Ces expressions sont de natures diverses, couvrant essentiellement des locutions et des expressions idiomatiques.

obtenues à partir d'un corpus de textes annotés manuellement. Une illustration de ce modèle HMM est donnée à la figure 3.

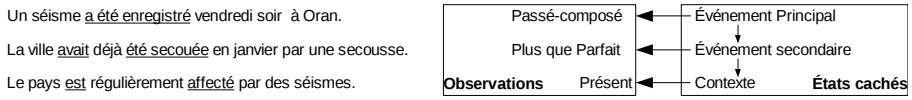

Figure 3. *Illustration de la segmentation de textes en événements avec le modèle HMM*

L'utilisation standard des HMM telle que décrite ci-dessus est limitée par le point suivant : pour une séquence d'observations donnée, le calcul de la séquence d'états correspondant ne considère que l'état précédent et l'observation courante. Il ne peut donc pas prendre pas en compte les dépendances existant avec les observations précédentes et ne permet pas de ce fait d'intégrer facilement une grande diversité de critères. Pour remédier à cette limitation, nous avons également défini un modèle fondé sur les CRF, décrit à la section suivante.

4.3. *Modèle CRF*

Depuis leur introduction en 2001, les CRF (Lafferty *et al.*, 2001) ont été, comme les HMM, très largement utilisés dans le domaine du TAL. Dans le cadre de la segmentation de textes avec catégorisation de segments, dont relève notre travail, Hirohata *et al.* (2008) ont ainsi obtenu de très bons résultats en appliquant un modèle CRF pour classifier les phrases contenues dans les résumés d'articles scientifiques selon quatre catégories : objectif, méthode, résultat, conclusion.

Les modèles HMM et les modèles CRF se différencient sur le point suivant : l'objectif des premiers est de maximiser la probabilité jointe $P(x, y)$ entre une séquence d'observations (x) et une séquence d'états cachés (y) alors que les seconds utilisent une approche conditionnelle (calcul de $P(y|x)$) pour attribuer une séquence d'états à une séquence d'observations. L'avantage de l'approche conditionnelle est de permettre la représentation de la séquence d'observations sous forme d'un vecteur dont les composantes sont issues de traits caractéristiques (ou *features*). Ces traits offrent la possibilité d'intégrer des connaissances variées dans les modèles. En reprenant Hirohata *et al.* (2008), une définition plus formelle des CRF est donnée par :

$$P(y|x) = \frac{1}{Z_\lambda(x)} exp(\lambda.F(y,x)) \qquad [1]$$

$$F(y,x) = \sum_i f(y,x,i) \qquad [2] \qquad Z_\lambda(x) = \sum_y exp(\lambda.F(y,x)) \qquad [3]$$

où $F(y, x)$ est un vecteur ayant pour composantes les valeurs des traits pour chaque élément i (ici une phrase) de la séquence d'entrée (ici un texte), λ est un vecteur

de pondération des traits, et $Z_\lambda(x)$ est un facteur de normalisation qui dépend de toutes les séquences d'états possibles. Un algorithme de programmation dynamique est généralement utilisé pour réduire la complexité du calcul de $Z_\lambda(x)$. De même que pour les modèles HMM, l'algorithme de Viterbi est utilisé pour calculer la séquence d'états la plus probable en fonction d'une séquence d'observations.

Nous intégrons les traits temporels suivants à notre modèle de segmentation événementielle :

– **le temps des verbes** : comme avec notre modèle HMM, nous faisons l'hypothèse que les changements de temps grammaticaux, en particulier lorsqu'ils concernent des temps du passé, sont corrélés aux changements d'événements dans le type de textes que nous considérons. Nous prenons en compte cette dimension dans notre modèle CRF en utilisant un trait binaire pour chaque temps grammatical possible, le trait valant 1 si au moins un verbe de la phrase est conjugué au temps considéré, 0 sinon ;

– **la présence d'une date** : si une phrase contient une date antérieure à la date de l'événement principal, il est probable qu'elle fasse référence à un événement secondaire. Nous exploitons cette caractéristique de façon limitée en utilisant un trait pour indiquer la présence ou l'absence d'une entité nommée de type date dans la phrase (dans le modèle actuel, la valeur de la date n'est pas utilisée) ;

– **les expressions temporelles** : ce trait est utilisé pour prendre en compte la présence d'une expression de localisation temporelle dans une phrase. Pour cela, nous utilisons un dictionnaire d'expressions que nous avons constitué manuellement à partir du corpus présenté dans (Laporte *et al.*, 2008). Le dictionnaire contient des expressions telles que : *au début de l'année, ces dernières années* ...

Par ailleurs, les dépendances de succession entre les différents statuts événementiels sont prises en compte au travers du caractère linéaire de notre modèle CRF. Un texte est en effet vu comme une suite de statuts événementiels.

5. Remplissage de formulaires événementiels

La seconde partie de notre processus d'extraction d'information se focalise sur les segments liés à l'événement principal mis en évidence par la segmentation événementielle présentée à la section précédente pour y rechercher les informations caractéristiques de cet événement et remplir le formulaire en faisant la synthèse. Pour cette recherche et ce remplissage, nous proposons une approche à base de graphe inspirée du paradigme de l'extraction de relations complexes. Bien que nous nous concentrions ici sur l'événement principal de chaque texte, cette approche n'est pas spécifique de ce type d'événements, en particulier du fait de la segmentation préalable des textes.

La première étape de cette approche identifie dans les textes les occurrences de toutes les entités susceptibles de remplir un champ du formulaire, occurrences que nous appellerons dans ce qui suit *mentions d'entités*. Cette identification inclut les *mentions d'événements*, c'est-à-dire les termes marquant la présence d'un événement

du type ciblé, qui correspond dans le formulaire du tableau 1 au champ particulier TYPE_EVT. La deuxième étape, dite de *construction du graphe*, détecte d'abord les relations existant au niveau d'une phrase entre deux mentions d'entités ou entre une mention d'entité et une mention d'événement, puis construit un graphe d'entités sur la base de la fusion des mentions d'événements et d'entités faisant référence à un même événement ou à une même entité. La dernière étape, dite de *remplissage du formulaire*, applique des stratégies génériques à ce graphe pour sélectionner les entités les plus à même de remplir le formulaire correspondant au type d'événement considéré. Ces étapes sont détaillées dans les trois sections suivantes.

5.1. *Identification des mentions d'événements et d'entités*

Le premier maillon de l'étape de remplissage d'un formulaire est double. Tout d'abord, il vise à identifier dans les textes, et plus précisément au sein des segments de texte liés à l'événement principal, les occurrences des éléments constitutifs de ce formulaire, c'est-à-dire les entités nommées dont le type est compatible avec les types d'entités associés aux différents champs du formulaire. Pour distinguer ces occurrences des entités elles-mêmes, nous appelons ces occurrences *mentions d'entités* : outre que chaque mention d'entité apparaît dans un contexte textuel particulier, les différentes mentions d'une même entité peuvent avoir des formes différentes lorsqu'un processus de normalisation permet de se ramener à une même entité, ce qui n'intervient dans notre cas que pour les entités numériques que sont les magnitudes, les dates et les heures. Dans le texte de la figure 4, la mention d'entité *7,2* est ainsi identifiée comme une magnitude possible du séisme correspondant à l'événement principal, de même que les mentions de lieux *Tokyo*, *Kurihara*, *préfecture de Miyagi*, *Iwate* ou *Miyagi* constituent des localisations possibles de ce séisme.

La seconde dimension de cette première étape est l'identification des mentions d'événements, c'est-à-dire des marques explicites dans les textes de la référence au type d'événement considéré. Cette tâche est aussi parfois appelée *détection d'événement*. En se focalisant sur les segments relatifs à l'événement principal, elle permet ainsi d'identifier dans le texte de la figure 4 les mentions d'événements *séisme*, *tremblement* et *secousse*, qui constituent trois façons de faire référence à un tremblement de terre dans un texte. Seules des formes nominales sont ici représentées – elles sont dominantes pour ce type d'événement – mais des formes verbales telles que « *La terre a tremblé dans les Hautes-Pyrénées ...* » peuvent aussi se rencontrer[2]. Dans notre approche, comme dans de nombreux travaux sur le remplissage de formulaires, la notion de mention d'événement est particulièrement importante car elle permet de repérer les entités les plus fortement liées aux événements considérés. Ce repérage s'appuie sur les indices existant au niveau phrastique, plus facilement exploitable en termes de moyens d'analyse que le niveau discursif. Au-delà, la mention d'événement occupe

2. Plus généralement, la fréquence relative des différentes formes de mentions d'événements peut varier selon le type des événements considérés.

une position de jointure entre ces deux niveaux : elle est en effet la manifestation au niveau phrastique d'une notion, celle d'événement, possédant une extension discursive, délimitée ici par la segmentation événementielle des textes.

Bien que conceptuellement distincts, les mentions d'entités et les mentions d'événements reposent sur les mêmes outils sur le plan de leur identification dans les textes. Celle-ci est en effet réalisée dans les deux cas par l'application d'un ensemble de règles élaborées manuellement et mises en œuvre grâce aux outils de l'analyseur LIMA sous la forme d'automates à états finis en s'appuyant sur les résultats d'une analyse morphosyntaxique des textes (désambiguïsation morphosyntaxique et lemmatisation).

5.2. *Construction du graphe d'entités*

La deuxième étape du processus de remplissage de formulaires a pour objectif de construire un graphe permettant de rassembler à l'échelle d'un document toutes les entités représentant des valeurs possibles des différents champs du formulaire associé à ce document. Afin de caractériser la saillance de ces entités, en vue de l'étape finale de sélection, il comprend également les relations extraites du document unissant ces entités les unes aux autres ou à l'événement sous-tendant le formulaire. La construction d'un tel graphe suppose donc deux opérations. La première extrait les relations identifiables au niveau des phrases entre tout couple de mentions d'entités ou entre une mention d'entité et une mention d'événement. La seconde réalise la fusion de toutes ces relations au niveau du document, passant ainsi de l'espace des occurrences textuelles à celui de la représentation consolidée d'un événement.

Le résultat est un graphe pondéré dont les nœuds représentent des entités ou des événements et les arcs, les relations qui les unissent. Ces relations étant symétriques, ce graphe est non dirigé. Un poids associé à chaque arc caractérise le degré de certitude de la présence d'une relation entre les deux entités liées. La figure 4 donne l'exemple d'un tel graphe pour le texte de la figure 1, graphe restreint aux entités liées à l'événement principal du document compte tenu de notre focalisation applicative[3]. Ce graphe fait en premier lieu apparaître que les différentes mentions d'entités ou d'événements sont rassemblées lorsqu'elles font référence à la même entité ou au même événement. Dans le cas présent, seul l'événement principal est concerné au travers des mentions *séisme*, *tremblement* et *secousse*. Cet exemple montre également que les relations détectées entre les mentions d'entités ou d'événements dans les textes se retrouvent sous la forme de relations entre les entités du graphe auxquelles elles se rattachent. La première phrase établit par exemple un lien entre une mention d'événement (*séisme*) et une mention de magnitude (*7,2*) qui se retrouve au niveau du graphe entre les entités

3. Pour être complet, il faut préciser que pour des raisons de lisibilité, le graphe ne fait pas apparaître l'entité *préfecture de Miyagi*, qui n'est pas associée à l'entité *Miyagi* dans la troisième phrase du fait de l'absence de normalisation des noms de lieux. Les relations de cette entité sont identiques à celles de l'entité *Kurihara*, à l'exception de leur poids.

correspondantes. À ce stade cependant, ces liens ne sont que des relations candidates fondées sur des critères linguistiques. À partir de ces critères, il est ainsi possible de proposer quatre localisations possibles du séisme principal mais le choix des localisations valides reste à faire.

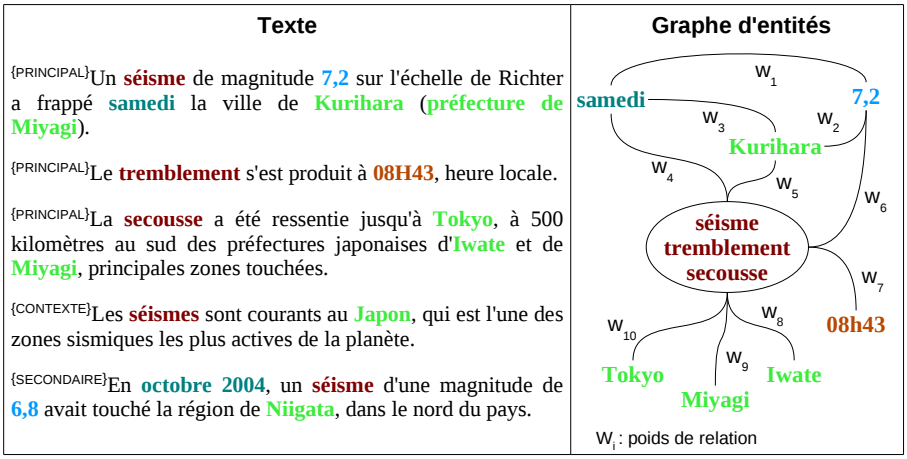

Figure 4. *Extrait du graphe d'entités pour l'exemple de la figure 1*

5.2.1. *Extraction des relations à l'échelle intraphrastique*

La construction du graphe d'entités d'un texte commence en déterminant si les couples de mentions d'entités ou d'événements apparaissant dans une même phrase sont sous-tendus par une relation propre au type d'événement cible, sans néanmoins préciser cette relation. À l'instar des travaux existants comparables, nous avons réalisé cette détermination par le biais d'un classifieur statistique. Dans ce cadre, l'utilisation d'un ensemble de traits lexicalisés constitue l'approche dominante (Afzal, 2009 ; Gu et Cercone, 2006 ; Wick *et al.*, 2006), même si Liu *et al.* (2007) se démarque en conjuguant ces traits lexicalisés avec des traits de nature syntaxique. À l'inverse, nous avons construit un modèle n'intégrant que des traits syntaxiques et faisant abstraction des informations lexicales (mots sous forme fléchie ou lemmes) afin de lui conférer un degré de généralité plus important susceptible de rendre son adaptation à un autre domaine plus facile. Pour évaluer l'intérêt relatif des traits syntaxiques et lexicaux, nous avons entraîné différents types de classifieurs avec les trois ensembles de traits suivants, détaillés dans le tableau 2[4] :

– *LEXI-BASE* : même ensemble de traits lexicalisés que Afzal (2009), à l'exception des bigrammes de mots ;

4. Certains traits apparaissant dans (Jean-Louis *et al.*, 2011a) ne figurent plus dans le tableau 2 car une étude plus approfondie a montré qu'ils n'étaient pas pris en compte par les classifieurs de MALLET

– *LEXI-SYN* : ensemble de traits conjuguant traits lexicalisés et traits syntaxiques, dans le prolongement de Liu *et al.* (2007)[5] ;

– *NON-LEXI-SYN* : même ensemble de traits que *LEXI-SYN*, à l'exception des traits lexicalisés.

Ces relations intervenant entre des couples de mentions d'entités ou d'événements, *E1* désigne dans le tableau 2 la première de ces mentions apparaissant dans la phrase et *E2*, la seconde. Outre leur nature intrinsèque (mot, catégorie morphosyntaxique ou type d'entité, relation de dépendance syntaxique ou caractéristique positionnelle), ces traits peuvent être différenciés en fonction de leur focalisation (séparation par une ligne double dans le tableau 2) : un premier ensemble s'attache aux entités mises en relation ; un deuxième à la partie du texte située entre les deux entités, supposée exprimer la relation, et un dernier ensemble, plus hétérogène, considère l'environnement immédiat des entités. Sur un plan plus technique, les traits donnés par le tableau 2 correspondent à des schémas de traits : chacun de ces schémas est en pratique décliné en autant de traits binaires que ce schéma admet de valeurs possibles. La construction de ces traits s'appuie comme dans le cas de la segmentation événementielle sur les résultats de l'analyseur linguistique LIMA dont nous exploitons en plus ici les capacités d'analyse syntaxique permettant de réaliser une analyse en dépendances syntaxiques des phrases.

Traits	LEXI-BASE	LEXI-SYN	NON-LEXI-SYN
Type d'entité de E1	✓	✓	✓
Type d'entité de E2	✓	✓	✓
Catégories morphosyntaxiques de E1	✓	✓	✓
Catégories morphosyntaxiques de E2	✓	✓	✓
Lemmes constitutifs de E1	✓		
Lemmes constitutifs de E2	✓		
Lemmes situés entre E1 et E2	✓	✓	
Catégories morphosyntaxiques situées entre E1 et E2	✓	✓	✓
Relations de dépendance syntaxique entre E1 et E2		✓	✓
Position/un événement[1]		✓	✓
Catégorie morphosyntaxique des deux mots avant/après E1		✓	✓
Catégorie morphosyntaxique des deux mots avant/après E2		✓	✓

[1] Si E1, respectivement E2, est une mention d'événement, ce trait donne la position de E2, respectivement E1, par rapport à elle (avant ou après).

Tableau 2. *Traits utilisés pour la classification de relations binaires*

Enfin, le poids associé à chaque relation trouvée correspond, comme dans (McDonald *et al.*, 2005) et (Liu *et al.*, 2007), au score de confiance du classifieur

5. Nous n'utilisons pas exactement les mêmes traits que Liu *et al.* (2007), en particulier parce que certains d'entre eux ne sont applicables que dans le domaine biomédical.

l'ayant mise en évidence, ce score étant compris dans l'intervalle [0,1] pour tous les classifieurs expérimentés à la section 6.4.

5.2.2. *Fusion des relations intraphrastiques*

La dernière étape de construction du graphe d'entités est une forme de condensation résultant de la fusion des mentions d'entités et d'événements identifiées comme faisant référence à une même entité ou à un même événement. Pour les événements, cette fusion s'appuie sur la segmentation événementielle : toutes les mentions d'événements apparaissant dans un segment étiqueté PRINCIPAL sont supposées faire référence à l'événement principal du document et sont donc fusionnées (cf. fusion de *secousse*, *séisme* et *tremblement* au niveau de la figure 4). Pour les entités, la fusion se fait sur l'égalité de leur forme normalisée dans le cas des dates, heures et magnitudes[6] et sur l'égalité de la forme trouvée dans les textes pour les lieux[7]. Lorsque l'opération de fusion entraîne la présence de plusieurs relations entre deux entités ou entre une entité et l'événement principal, ces relations sont elles-mêmes fusionnées en conservant le poids le plus élevé.

5.3. *Remplissage du formulaire*

L'étape de remplissage du formulaire a pour objectif de choisir pour chaque rôle de ce formulaire l'entité du graphe construit à l'étape précédente ayant un type compatible avec le type d'entité attendu pour ce rôle et se montrant la plus à même de le remplir. Cette sélection s'accompagne implicitement du choix de ne pas remplir certains rôles du formulaire lorsque les informations correspondantes sont absentes du texte. Ce problème de remplissage du formulaire peut être assimilé au problème de la reconstruction d'une relation complexe tel qu'il est envisagé dans (Afzal, 2009 ; McDonald *et al.*, 2005). Par exemple, le graphe de la figure 4 comporte une ambiguïté relative à l'entité occupant le rôle de lieu de l'événement et impose un choix entre : *Kurihara*, *Tokyo*, *Miyagi* ou *Iwate*. Dans cette perspective, nous avons testé plusieurs approches :

Position est une heuristique simple mais très efficace dans le contexte considéré qui sélectionne pour chaque type d'entités la première mention apparaissant dans un segment relatif à l'événement principal ;

Confiance retient pour chaque type d'entités l'entité liée à l'événement avec le score de confiance (score du classifieur utilisé) le plus grand ;

PageRank est une approche exploitant la structure globale du graphe d'entités par le biais de l'algorithme PageRank. Ce dernier permet en l'occurrence d'attribuer un score d'importance à chaque entité en fonction de sa connectivité avec

6. Par exemple pour les dates, des mentions telles que *[samedi, 14 juin, 14/06/08]* sont fusionnées puisqu'elles ont la même valeur normalisée *14/06/2008*.

7. Les entités de type DOMMAGES ont été laissées de côté ici car la variabilité de leur expression rend leur appariement complexe, proche de la problématique de la paraphrase.

les autres entités et donc de les ordonner. Pour chaque type d'entités, est ainsi retenue l'entité ayant le plus haut score PageRank ;

Vote implémente une stratégie de vote majoritaire reposant sur les approches *Position*, *Confiance* et *PageRank*. Pour chaque type d'entités, l'entité ayant été sélectionnée par le plus grand nombre d'approches est ainsi adoptée ;

Hybride applique pour chaque type d'entités celle, parmi les stratégies précédentes, donnant le meilleur résultat pour ce type d'entités.

La sortie des approches *Confiance*, *PageRank*, *Vote* et *Hybride* est en outre complétée par l'approche *Position* dans le cas où aucune entité n'est sélectionnée pour un type donné. Il est en effet possible que certaines entités d'un formulaire apparaissent dans un texte sans être associées dans une phrase à une mention d'entité ou d'événement, ce qui interdit leur choix par les approches reposant sur le graphe d'entités.

6. Évaluation

Nous présentons dans cette section une évaluation de notre approche de remplissage de formulaires sur un corpus de dépêches de presse concernant les événements sismiques, que nous décrivons à la section 6.1. Une évaluation différenciée de chaque étape de notre approche a été menée : la segmentation événementielle à la section 6.2, la construction du graphe d'entités et la sélection des entités aux sections 6.4 et 6.5, sans oublier une évaluation de la qualité de la reconnaissance des entités nommées impliquées dans les formulaires à la section 6.3. Une évaluation plus ciblée de l'impact de la segmentation en événements sur le résultat final est présentée à la section 6.6 et une analyse des principales erreurs rencontrées et de leur répartition est présentée dans la section 6.7.

6.1. *Corpus*

Les travaux présentés dans cet article ont été développés dans le cadre d'une application dédiée à la surveillance des événements sismiques à partir de dépêches de presse. Dans ce contexte, un formulaire est associé à un événement sismique et résume ses principales caractéristiques telles qu'elles sont décrites dans le tableau 1.

L'évaluation des modèles développés pour le remplissage de ces formulaires a été effectuée à partir d'un corpus composé de 501 dépêches de presse en français concernant le domaine sismique. Ces dépêches ont été collectées entre fin février 2008 et début septembre 2008, en provenance pour partie d'un flux de dépêches AFP (un tiers du corpus), et pour partie de dépêches collectées à partir de Google Actualités (deux tiers du corpus). Le corpus contient à la fois des dépêches ayant une structure simple (un seul événement) et une structure complexe (plusieurs événements) : 252 dépêches (50 %) mentionnent au moins un événement secondaire. Le corpus a été annoté par des analystes du domaine qui ont rempli manuellement les formulaires pour chaque

séisme principal d'un document. Au total, les annotateurs ont ainsi identifié 3 306 entités se répartissant entre sept types (incluant les mentions d'événements) comme indiqué dans le tableau 3. La possibilité a été laissée aux annotateurs de retenir plus d'une entité pour le même rôle lorsque plusieurs variantes étaient mentionnées et étaient jugées également pertinentes. Par exemple, pour les lieux, pouvaient être annotés à la fois un nom de ville et un nom de pays. On peut remarquer que la distribution des entités nommées n'est pas uniforme : on trouve ainsi beaucoup de noms de lieux (*28,6 %*) mais très peu de coordonnées géographiques (*0,9 %*). Du point de vue de la complétude des formulaires, ce dernier point peut être compensé par le recours à des bases de connaissances géographiques, à l'image de Gazetiki (Popescu *et al.*, 2008), qui permettent d'avoir accès aux coordonnées géographiques de lieux nommés, tels que des villes par exemple. Cette dimension n'a cependant pas été abordée ici.

Type d'entité	Nombre	Pourcentage
TYPE_EVT	499	15,1
LIEU	947	28,6
DATE	470	14,2
HEURE	345	10,4
MAGNITUDE	484	14,6
COORD_GEO	30	0,9
DOMMAGES	531	16,1

Tableau 3. *Distribution des entités nommées dans le corpus de référence*

Comme indiqué précédemment, ce corpus a fait l'objet d'un prétraitement par l'analyseur linguistique LIMA allant jusqu'à l'analyse syntaxique.

6.2. *Segmentation des textes en événements*

Pour évaluer notre approche de segmentation en événements, nous avons annoté manuellement en segments événementiels une sous-partie de notre corpus de référence composée de 140 dépêches, principalement sélectionnées parmi les dépêches évoquant au moins un événement secondaire. Le tableau 4 montre la distribution des événements sur la sous-partie annotée. On remarque que la catégorie d'événements la plus représentée est *Événement principal (70 %)*, ce qui est cohérent avec l'aspect très factuel des dépêches de presse. La catégorie *Événement secondaire* regroupe sans distinction tous les événements différents de l'événement principal : notons que parmi les dépêches sélectionnées, le nombre réel d'événements secondaires différents évoqués peut monter jusqu'à quatre, avec un nombre moyen de 1,66 événements secondaires évoqués.

Les résultats de l'évaluation de nos deux modèles de segmentation événementielle

1 659 phrases dans 140 dépêches		
Statut événementiel	Nombre	Représentativité
Événement principal	1 168	70 %
Événement secondaire	287	17 %
Contexte	213	13 %

Tableau 4. *Distribution des statuts événementiels dans la sous-partie du corpus de référence utilisée pour l'évaluation de la segmentation événementielle*

(*HMM* et *CRF*[8]) sur ce corpus de 140 dépêches sont donnés dans le tableau 5. Ces résultats sont exprimés en termes de précision et de rappel (notées P et R) en se plaçant dans l'optique d'une tâche de classification de phrases. Compte tenu de la petitesse du corpus d'évaluation, ils ont été obtenus par validation croisée avec un découpage en cinq parties. Pour comparaison, nous présentons également dans ce tableau les résultats des trois autres méthodes suivantes :

– *ParaSeg* : cette *baseline* attribue la catégorie *Principal* à toutes les phrases des deux premiers paragraphes et considère les autres phrases comme étant associées à la catégorie *Secondaire* ;

– *HeurSeg* : cette méthode heuristique est également considérée comme une *baseline* mais utilise comme critère principal la présence et la valeur des dates selon les principes suivants : des dates ayant des valeurs différentes correspondent à des segments différents (le segment principal étant celui de date la plus récente) ; les ruptures de segments entre deux dates différentes s'appuient sur la structure du texte en phrases et paragraphes ainsi que sur la présence d'autres entités caractéristiques du domaine entre les dates ;

– *MaxEnt* : pour juger plus clairement de l'intérêt du modèle CRF, nous avons entraîné un modèle de type *Maximum d'Entropie*[9] avec les mêmes traits que le modèle CRF, ce qui est équivalent à un modèle CRF sans dépendance entre les états.

En se référant au tableau 5, on peut noter en premier lieu que les baselines *ParaSeg* et *HeurSeg* sont toutes deux dépassées en termes de précision par une heuristique assignant toutes les phrases à l'événement principal. Par ailleurs, si les résultats de *ParaSeg* et *HeurSeg* sont comparables du point de vue de la précision, le rappel de *ParaSeg* est véritablement très faible. Le fait que la plupart des articles sont constitués de courts paragraphes et que donc, seul un petit nombre de phrases peuvent être rattachées à l'événement principal dans le cas de *ParaSeg* explique ce faible rappel. On notera également l'absence de prise en compte de la catégorie *Contexte* par cette heuristique.

8. Implémentés respectivement grâce aux outils NLTK (`http://www.nltk.org/`) et CRF++ (`http://crfpp.sourceforge.net/`).

9. Implémenté grâce à l'outil développé par Dekang Lin (`http://webdocs.cs.ualberta.ca/~lindek/maxent.tgz`).

	ParaSeg		HeurSeg		HMM		MaxEnt		CRF	
Type d'événement	R	P	R	P	R	P	R	P	R	P
Événement principal	6,1	63,9	82,8	64,7	83,0	93,6	94,8	78,7	98,7	87,4
Événement secondaire	86,9	12,4	23,5	43,4	37,8	9,6	33,6	54,7	52,6	95,8
Contexte	–	–	16,9	21,7	49,1	40,0	22,0	84,2	69,3	93,0

Tableau 5. *Résultats de la segmentation des textes en événements (en %)*

Concernant les modèles de segmentation statistiques, les résultats du tableau 5 montrent au travers du modèle HMM que le seul critère de la succession des temps grammaticaux n'est pas suffisant pour discriminer tous les types d'événements : tandis que l'*Événement principal* est correctement identifié (83,0 % de rappel et 93,6 % de précision), les résultats pour les catégories *Événement secondaire* et *Contexte* sont nettement moins bons. Cette insuffisance n'est pas gênante dans notre contexte particulier mais pourrait l'être dans un contexte plus général. Les modèles MaxEnt et CRF apportent de ce point de vue une amélioration significative, avec un avantage au modèle CRF confirmant l'intérêt de prendre en compte la succession des types d'événements au niveau textuel. C'est donc le modèle CRF qui a été utilisé par la suite pour l'évaluation globale du processus d'extraction d'information. Il faut néanmoins noter que même pour le modèle CRF, la valeur du rappel pour les événements secondaires reste assez faible, ce qui est aussi vrai pour la catégorie *Contexte* dans une moindre mesure. Cette faiblesse s'explique avant tout par le fort déséquilibre du corpus considéré en faveur de la catégorie *Événement principal*, comme nous l'avons vu avec le tableau 4 : le classifieur optimise naturellement sa décision en fonction de la classe majoritaire. Ceci se traduit d'ailleurs par une valeur de précision moindre pour cette classe. Compte tenu de l'effectif de la catégorie *Événement principal* par rapport aux deux autres, le classifieur peut en effet favoriser cette catégorie sans faire chuter la précision de façon importante.

En l'absence de travaux fournissant des résultats directement comparables aux nôtres, seuls les résultats du *Relevant Sentence Classifier* de Patwardhan et Riloff (2007) peuvent être donnés à titre d'éléments indicatifs, avec des valeurs de rappel|précision égales à 63 %|46 % pour le domaine du terrorisme (corpus MUC-4) et à 72 %|41 % pour celui des maladies (corpus ProMed). Il faut préciser néanmoins que cette méthode, qui s'appuie sur un classifieur de type SVM exploitant des traits lexicalisés et non des indices temporels, juge simplement si une phrase est pertinente vis-à-vis des informations recherchées et que son évaluation a été menée pour l'anglais dans des domaines assez éloignés du domaine sismique.

	Complet_50			Partiel_500
Type d'entité	Précision	Rappel	F1-mesure	Rappel
TYPE_EVT	93,9	93,0	93,4	97,4
LIEU	90,5	66,5	76,6	84,4
DATE	88,2	86,3	87,2	98,7
HEURE	82,6	86,5	84,5	96,5
MAGNITUDE	93,8	83,3	88,2	94,0
DOMMAGES	83,5	63,9	72,4	62,7
COORD_GEO	100,0	66,7	80,0	86,7
toutes entités	89,8	77,4	83,2	72,9

Tableau 6. *Évaluation de la reconnaissance des entités nommées et des événements sur un corpus de 50 dépêches annotées complètement et sur un corpus de 501 dépêches annotées partiellement (en %)*

6.3. *Reconnaissance des entités nommées et détection des événements*

L'évaluation de la reconnaissance des entités nommées impliquées dans la description des événements sismiques, incluant la détection des événements au travers de l'identification de leurs mentions, a été menée sur deux corpus : un premier corpus de taille réduite, 50 dépêches, mais annoté de façon exhaustive pour les entités considérées ; un second corpus de plus grande taille, 501 dépêches de notre corpus de référence, mais avec une annotation partielle des entités. Dans ce dernier cas, seules les entités liées à l'événement principal ont en effet été annotées. Pour cette seconde évaluation, seul le rappel est donné car les entités n'étant pas toutes annotées dans la référence, la précision n'aurait pas vraiment de sens : une entité correctement identifiée par LIMA serait en effet comptée comme fausse si elle ne fait pas partie des entités annotées manuellement.

Les résultats de ces deux évaluations sont présentés dans le tableau 6, avec des mesures données sous forme de pourcentages. Même si ces résultats sont plutôt inférieurs à l'état de l'art pour certaines entités largement répandues (lieux par exemple), les performances sont acceptables. En particulier, le taux de rappel est très bon sur le corpus de 501 dépêches, ce qui est encourageant pour les traitements postérieurs qui s'appuient sur ces résultats. Les moins bons scores obtenus pour un type d'entité comme DOMMAGES s'expliquent par une plus grande variabilité d'expression des entités : les annotations de référence comptent ainsi des expressions simples comme « *65 blessés* » et des expressions plus complexes comme « *faisant six morts, interrompant la circulation des trains, provoquant des glissements de terrain et l'effondrement d'un pont* ».

[POSITIVE] : Cette *secousse*, d'une magnitude de *6,4* sur l'échelle de Richter, est la plus forte enregistrée depuis le tremblement de terre d'une magnitude de 8 qui a ravagé le Sichuan, a précisé un responsable du bureau de sismologie de cette province.

[NEGATIVE] : Cette *secousse*, d'une magnitude de 6,4 sur l'échelle de Richter, est la plus forte enregistrée depuis le tremblement de terre d'une magnitude de *8* qui a ravagé le Sichuan, a précisé un responsable du bureau de sismologie de cette province.

Figure 5. *Exemples positif et négatif de présence d'une relation entre deux entités*

6.4. *Construction du graphe d'entités*

La méthode proposée pour la construction du graphe d'entités s'appuie sur un classifieur pour déterminer la présence ou l'absence d'une relation entre deux entités au sein d'une même phrase. Nous avons expérimenté différents types de classifieurs statistiques en testant pour chacun d'entre eux les trois ensembles de traits présentés à la section 5.2 (*LEXI-BASE, LEXI-SYN, NON-LEXI-SYN*). Pour l'annotation des relations binaires entre entités, nous avons considéré un sous-ensemble du corpus composé de 44 dépêches. À partir de ce sous-ensemble, nous avons obtenu 5 000 relations binaires, parmi lesquelles 969 relations sont exprimées à l'intérieur de la même phrase. Parmi celles-ci, 43 relations ont été écartées à cause d'erreurs de reconnaissance des entités (par exemple, lorsqu'une entité considérée est en fait incluse dans une entité plus large non reconnue à cause de son type). Les autres relations ont servi pour l'entraînement des classifieurs : 690 pour la catégorie *POSITIVE*, dans laquelle les deux entités font référence au même événement sismique et 236 pour la catégorie *NEGATIVE*, dans laquelle les deux entités sont associées à des événements sismiques différents. La figure 5 illustre des relations pour les deux catégories.

Ce corpus annoté nous a servi à tester trois types de classifieurs[10] : bayésien naïf (*NB*), maximum d'entropie (*ME*) et arbres de décision (*DT*). Nous reportons dans le tableau 7 les résultats produits par chaque algorithme, en fonction de l'ensemble de traits utilisé, en termes de rappel (*R*), précision (*P*) et F1-mesure (*F*). En complément, nous fournissons pour comparaison les résultats d'une approche basique (*Baseline*) qui attribue la catégorie la plus fréquente (*POSITIVE*) à toutes les relations. Pour les classifieurs, les résultats sont obtenus par une procédure de segmentation aléatoire itérative du corpus d'évaluation entre données d'entraînement et données de test : 4/5 des données servent à l'entraînement et 1/5 pour le test à chaque itération selon une sélection aléatoire. Les résultats sont moyennés sur cinq itérations. Contrairement à une validation croisée classique (qui n'était pas proposée par la version de MALLET que nous avons utilisée), cette procédure ne garantit pas l'utilisation de chaque exemple à

10. Nous avons utilisé les implémentations fournies par l'outil MALLET.

la fois pour l'entraînement et pour le test mais nous n'avons pas observé en pratique de différence avec une véritable validation croisée.

Ensemble de traits	Classifieur	R(%)	P(%)	F(%)
LEXI-SYN	ME	95,9	96,3	96,1
LEXI-BASE	ME	96,1	91,2	93,6
NON-LEXI-SYN	ME	95,0	91,7	93,3
LEXI-SYN	DT	96,5	89,0	92,5
LEXI-SYN	NB	90,7	93,4	92,0
NON-LEXI-SYN	DT	88,7	91,2	89,8
NON-LEXI-SYN	NB	89,2	89,6	89,4
LEXI-BASE	DT	94,7	84,3	89,2
LEXI-BASE	NB	87,9	86,7	87,3
Baseline	–	100,0	25,5	40,5

Tableau 7. *Évaluation du classifieur de relations binaires*[11]

Du point de vue des ensembles de traits considérés, les résultats du tableau 7 montrent d'abord l'intérêt d'utiliser des traits de nature syntaxique : les scores obtenus à partir de l'ensemble LEXI-SYN dépassent ceux de l'ensemble LEXI-BASE pour les trois classifieurs. Ils montrent également que l'ensemble de traits non lexicalisés NON-LEXI-SYN obtient des scores assez comparables à ceux de l'ensemble LEXI-BASE : des modèles non lexicalisés, *a priori* plus génériques, ne sont donc pas nécessairement pénalisés par rapport à des modèles fondés sur des traits plus spécifiques à un domaine. Concernant les algorithmes d'apprentissage, le point marquant est la supériorité du classifieur ME par rapport aux deux autres. Ces deux derniers peuvent être considérés comme équivalents compte tenu de la faible différence au niveau des résultats. Notons que Afzal (2009) obtient une hiérarchie différente (DT $>$ ME $>$ NB) mais utilise un corpus différent et dans une autre langue, ce qui rend la comparaison difficile. En termes de performances générales, nos résultats sont comparables à ceux présentés par Afzal (2009), ses meilleurs scores étant R = 95 %|P = 87 %|F = 91 %, obtenus avec des arbres de décision. Pour la suite de notre démarche, nous avons conservé le classifieur de type maximum d'entropie reposant sur l'ensemble de traits NON-LEXI-SYN plutôt que l'ensemble LEXI-SYN. Notre motivation pour ce faire est que l'ensemble NON-LEXI-SYN permet d'obtenir des scores satisfaisants sans être fondé sur des informations fortement liées à un domaine, ce qui n'est pas le cas pour les traits lexicalisés.

11. On notera que par rapport aux résultats publiés dans (Jean-Louis *et al.*, 2011a), les chiffres de la précision et du rappel sont inversés pour tous les couples (ensemble de traits, classifieur) en dehors de la condition *Baseline*. Une erreur d'interprétation de la table de contingence en sortie de MALLET a en effet conduit à intervertir la précision et le rappel dans cet article.

6.5. *Sélection des entités et remplissage des formulaires*

Concernant l'évaluation des stratégies de sélection, l'ensemble des documents du corpus a été utilisé. Nous reportons dans le tableau 8 les scores de remplissage des formulaires pour ces différentes stratégies, agrégés pour l'ensemble des rôles du formulaire, en termes de rappel (*R*), précision (*P*) et F1-mesure (*F*).

Stratégie de sélection	**R(%)**	**P(%)**	**F(%)**
Hybride	77,5	76,9	77,2
Vote	74,9	74,3	74,5
Confiance	74,9	74,2	74,5
Position	73,4	73,1	73,2
PageRank	72,4	71,7	72,0

Tableau 8. *Évaluation du remplissage des formulaires selon les stratégies de sélection*

Ces résultats confirment en premier lieu que notre méthode de référence *Position* est caractérisée par un niveau déjà très élevé. De plus, cette méthode permet d'obtenir des performances légèrement supérieures à la stratégie *PageRank*, ce qui peut se justifier en partie par le fait que la stratégie *PageRank* repose uniquement sur la structure du graphe, sans tenir compte des poids sur les arcs. Par conséquent, les entités se trouvant dans des zones densément connectées du graphe obtiennent de meilleurs scores que les autres, indépendamment des poids sur les arcs. Ce problème pourrait être, dans une certaine mesure, minimisé en adoptant la version pondérée de l'algorithme PageRank proposée dans (Mihalcea, 2004). D'autre part, les scores du tableau 8 montrent que la meilleure stratégie de sélection est l'approche *Hybride*, ce qui est cohérent avec ses objectifs de faire correspondre à un rôle du formulaire la stratégie qui lui est la mieux adaptée.

6.6. *Impact de la segmentation en événements*

Dans cette section, nous proposons d'évaluer l'impact de notre approche de segmentation en événements sur la tâche de remplissage des formulaires. Cette segmentation vise à identifier les passages pertinents afin de focaliser le processus d'extraction. Cependant, tous les documents ne mentionnent pas plusieurs événements sismiques et dans le cas des documents ne mentionnant qu'un seul événement, l'usage de la segmentation événementielle se justifie moins car toutes les phrases comportant une mention d'événement font *a priori* référence au même événement. Son intérêt dans une telle situation n'est en principe pas nul car elle permet alors d'éliminer les phrases relatives au contexte. En pratique, la segmentation est plutôt vue comme une source potentielle de perturbations pour les documents mono-événements car l'intérêt de la détection des phrases de contexte, très minoritaires en nombre, est à mettre en balance avec des erreurs de segmentation pouvant conduire à des baisses de rappel.

Notre but, dans cette section, est donc de mesurer l'impact de la segmentation en événements sur les documents ne faisant référence qu'à un seul événement en comparaison avec ceux faisant référence à plusieurs événements. Notre intuition est que la segmentation devrait avoir un impact limité sur les documents mono-événements et devrait améliorer les scores pour les documents multi-événements. Afin de vérifier cette hypothèse, nous avons manuellement divisé le corpus initial en deux ensembles en fonction du nombre d'événements sismiques mentionnés par les textes. Nous avons ainsi obtenu 227 documents multi-événements (*M*) et 274 documents mono-événements (*S*). Les résultats du remplissage de formulaires pour chaque ensemble, avec ou sans segmentation, sont présentés dans le tableau 9 en termes de F1-mesure et agrégés pour l'ensemble des rôles du formulaire.

	Sans segmentation		**Avec segmentation**	
Stratégie	S (%)	M (%)	S (%)	M (%)
Hybride	79,2	73,6	78,3	75,6
Vote	77,7	68,7	76,9	71,8
Confiance	72,5	66,1	71,8	69,1
Position	74,0	73,2	73,1	73,1
PageRank	70,9	59,7	70,7	65,3

Tableau 9. *Impact de la segmentation sur les documents mono et multi-événements (F1-mesure)*

Concernant les documents mono-événements, les scores du tableau 9 montrent que les stratégies les plus performantes n'utilisent pas de segmentation, bien que les différences ne soient pas très importantes (+ 0,71 % en moyenne). À l'opposé, les stratégies à base de segmentation sont plus performantes pour les documents multi-événements (+ 2,74 % en moyenne). De plus, notre stratégie la plus performante, approche *Hybride* avec segmentation, obtient de meilleurs scores que notre approche de référence, *Position* sans segmentation, et ce, pour les deux ensembles de documents. Plus généralement, les résultats démontrent que notre segmentation n'introduit qu'une perte limitée pour les documents mono-événements et améliore les performances pour les documents multi-événements.

6.7. *Analyse des erreurs*

Afin d'approfondir les évaluations de nos stratégies de remplissage de formulaires, nous avons mené une analyse des erreurs en cherchant à identifier précisément les causes de la présence d'une entité incorrecte (sélection d'une mauvaise entité pour un rôle) ou d'une entité manquante (pas d'entité sélectionnée pour un rôle) dans un formulaire. Dans ce cadre nous avons identifié trois types prépondérants d'erreurs :

– les erreurs de reconnaissance des entités nommées : l'entité n'est pas reconnue lors de l'analyse linguistique du texte ;

– les erreurs de segmentation en événements : l'entité est identifiée lors de l'analyse linguistique mais elle appartient à une phrase qui n'est pas associée à l'événement principal ;

– les erreurs de sélection des entités : l'entité se trouve dans le segment de l'événement principal mais une autre entité a été retenue comme valeur pour le rôle dans le formulaire.

Le tableau 10 présente la répartition de chaque type d'erreurs, en comparaison avec le nombre d'entités correctement repérées, pour deux approches de construction des formulaires : la première correspond à la sélection à base d'heuristique, sans segmentation (*NonSeg+Position*) ; la seconde s'appuie sur la segmentation en événements et la stratégie *Seg+Hybride*. Le tableau montre que la stratégie de référence *Position*

		NonSeg+Position	Seg+Hybride
	Correct	71,6 %	75,1 %
Type d'erreurs	Sélection d'entités	25,6 %	21,2 %
	Reconnaissance d'entités	2,8 %	2,8 %
	Segmentation	–	0,8 %

Tableau 10. *Répartition des erreurs pour le remplissage de formulaires*

permet d'identifier correctement une part conséquente des entités (*71,6 %*) mais qu'un nombre important d'erreurs d'attribution de rôle dans le formulaire (*25,6 %*) subsiste. Notre meilleure stratégie réduit ce type d'erreurs tout en améliorant le pourcentage d'entités correctes dans les formulaires. De plus, cette stratégie n'induit qu'un nombre très limité d'erreurs dues à la segmentation en événements (*0,8 %*).

Une analyse plus fine des résultats par type d'entités fait apparaître que pour des résultats moyens de 77,2 % (F1-mesure pour notre meilleure stratégie), les valeurs vont de 58,7 % pour les entités LIEU jusqu'à 92,9 % pour les entités COORD_GEO. Plus précisément, les entités LIEU, DATE et dans une moindre mesure HEURE obtiennent les scores de rattachement les plus faibles, ce qui peut s'expliquer par au moins deux facteurs. En premier lieu, les textes considérés contiennent de façon générale plus d'entités LIEU que d'entités COORD_GEO ou même MAGNITUDE par exemple (cf. tableau 3), ce qui rend intrinsèquement le choix plus difficile pour ce type d'entités. Ensuite, la forme d'expression des relations entre les mentions d'événements et d'entités est plus ou moins variée et plus ou moins difficile à analyser selon les types d'entités : une MAGNITUDE sera presque toujours précédée du nom *magnitude*, ce qui fige un peu les expressions utilisées pour l'évoquer, alors qu'un LIEU apparaît souvent sous la forme d'un complément circonstanciel, plus difficile à rattacher dans une analyse syntaxique. De ce point de vue, particulariser les modèles d'extraction de relations intraphrastiques en fonction du type des entités ou plus précisément de leur rôle vis-à-vis du formulaire pourrait être une voie d'amélioration.

7. Discussion et positionnement par rapport aux travaux comparables

Une des spécificités du travail que nous avons présenté dans cet article est son utilisation d'une segmentation événementielle, segmentation reposant sur des indices de nature temporelle. L'usage de tels indices pour la segmentation du discours a principalement été exploré d'un point de vue linguistique et psycholinguistique au travers de l'étude du rôle en tant que marqueurs de segmentation des adverbes temporels en initial de proposition. Sur le plan psycholinguistique, Bestgen et Vonk (2000) ont ainsi montré que les changements thématiques sont corrélés avec la présence d'adverbes temporels en tête de proposition. Sur le plan linguistique en revanche, Ho-Dac et Péry-Woodley (2008) présentent une situation plus complexe dans laquelle ces adverbiaux ont un rôle en tant que marqueurs de segmentation variant selon les types de discours. Notre usage des indices temporels est à la fois plus large et moins différencié : nous nous appuyons principalement sur la séquence des temps grammaticaux et nous utilisons les autres indices temporels, comme la présence de dates ou d'adverbes temporels, pour différencier plus finement les événements secondaires et les segments de contexte.

La segmentation des textes du point de vue de la dimension événementielle a, quant à elle, été abordée par quelques travaux. Dans (Kitani *et al.*, 1994), elle est principalement fondée sur l'identification dans les textes des constituants caractéristiques des types d'événements considérés, ces constituants provenant d'une connaissance *a priori* sur le domaine. Néanmoins, elle repose également sur deux structures discursives caractéristiques des articles de presse traités : l'une prend la forme d'une séquence d'événements tandis que la seconde est structurée, comme dans notre cas, autour d'un événement principal et de références à des événements secondaires. Crowe (1995) fonde également sa segmentation des textes sur l'identification des constituants des événements pour une part importante mais utilise simultanément plusieurs heuristiques d'inspiration discursive pour rattacher une proposition à un événement, comme par exemple le fait de favoriser l'événement du rattachement le plus récent. Finalement, le travail le plus proche du nôtre, du point de vue de la segmentation, est celui présenté dans (Naughton, 2007), travail dans lequel les phrases sont étiquetées selon quatre statuts événementiels : nouvel événement, continuation d'événement, référence à un événement précédent, sans événement. Ces statuts sont proches de nos trois types d'événements en dehors de l'absence de différenciation de l'introduction d'un nouvel événement et de sa continuation. Le modèle défini dans (Naughton, 2007) est un automate à états finis probabiliste reposant sur l'algorithme MDI pour sa partie apprentissage. Bien que l'objectif de ce travail soit de modéliser la structure discursive des textes sur le plan événementiel, il diffère du nôtre dans sa volonté de modéliser directement la séquence des statuts événementiels, sans ancrage dans une information temporelle, ce qui laisse craindre *a priori* une plus grande dépendance vis-à-vis d'un type de discours particulier. Naughton (2007) montre l'impact positif de sa segmentation sur sa tâche finale, en l'occurrence le regroupement de phrases référant à un même événement dans des dépêches de presse, mais ne donne pas d'évaluation directe de cette segmentation.

Si l'on considère plus globalement le processus d'extraction d'information exposé dans cet article, les connexions entre notre approche et les travaux existants interviennent à plusieurs niveaux. La stratégie générale consistant à d'abord identifier au niveau textuel les zones présentant des informations intéressantes pour ensuite les extraire se retrouve notamment dans (Patwardhan et Riloff, 2007), (Gu et Cercone, 2006) ou dans (Warnier et Nédellec, 2011), ces derniers ayant une approche « en creux » puisqu'ils éliminent les phrases jugées non porteuses d'information. Notre spécificité se situe ici dans le degré de généricité de ce processus. Dans (Patwardhan et Riloff, 2007), (Gu et Cercone, 2006) ou (Warnier et Nédellec, 2011), il est complètement lié au domaine considéré et aux informations recherchées alors que notre segmentation événementielle, même si elle n'est pas complètement dégagée de la tâche à laquelle elle contribue, présente un caractère de généricité plus grand, obtenu en s'appuyant sur des indices temporels « généraux ».

Au niveau du remplissage des formulaires proprement dit, la stratégie d'association d'une extraction de relations locales, en général au niveau phrastique, et de la construction d'un graphe d'entités pour synthétiser les relations entre entités au niveau textuel se retrouve de façon plus ou moins explicite dans divers travaux. L'extraction des relations locales est fréquemment réalisée par le biais de classifieurs statistiques qui, à la suite de McDonald *et al.* (2005),[12] peuvent s'appuyer à ce niveau sur un ensemble assez riche de traits allant jusqu'aux relations syntaxiques. La construction et l'exploitation du graphe de relations issu de cette extraction donnent lieu à des approches plus diverses. Afzal (2009) construit un graphe assez similaire au nôtre mais opère le remplissage des formulaires en recherchant des cliques maximales. Il obtient ainsi une F1-mesure de 64,6 % sur les données issues de MUC-6 (Grishman et Sundheim, 1996). Bien que nos données d'évaluation ne soient pas celles de MUC-6, la différence de résultat suggère qu'au-delà de l'exploitation structurelle du graphe d'entités réalisée par Afzal (2009), la prise en compte de la qualité de l'identification des relations locales peut apporter un plus.

Dans le domaine biomédical, source de beaucoup de travaux récents en extraction d'information au travers de la *BioNLP Shared Task*, l'approche initialement la plus performante (Björne *et al.*, 2009) reprenait un schéma assez similaire, avec néanmoins un remplissage des formulaires à partir du graphe d'entités fondé sur une série d'heuristiques propres au domaine et aux types d'événements considérés. Miwa *et al.* (2010) a globalement adopté la même philosophie en structurant plus explicitement cette tâche de remplissage et en substituant des méthodes à base d'apprentissage aux heuristiques, notamment pour résoudre le problème du rattachement d'un événement à un autre événement[13]. Plus récemment, Riedel et McCallum (2011) ont proposé une approche plus radicalement différente et plus intégrée en développant des mo-

12. Il est à noter que dans le cas de McDonald *et al.* (2005), le graphe des relations construit se limite à l'espace de la phrase.

13. À la différence des événements sismiques que nous considérons, certains événements de la *BioNLP Shared Task*, comme par exemple les événements *Regulation*, peuvent avoir comme arguments d'autres événements.

dèles probabilistes capables de réaliser de façon conjointe plusieurs tâches, comme la détection des événements et le remplissage des formulaires ou le remplissage de formulaires et la prise en compte de contraintes entre arguments d'événements liés. Cette problématique des modèles joints se retrouve dans (Reichart et Barzilay, 2012) pour intégrer les résultats des tâches d'identification des événements et des entités pertinentes tout en prenant en compte un ensemble de contraintes relatives à la fois à l'agencement des événements du point de vue discursif et à la cohérence des formulaires construits pour les représenter. Ces modèles présentent l'avantage de capturer les interactions existant entre les différentes tâches du processus d'extraction d'information mais restent encore très exigeants du point de vue calculatoire. Ce type d'approche pourrait néanmoins être intéressant à adapter au niveau de notre processus final de sélection des entités afin d'exploiter de façon plus spécifique, que ne peut le faire un algorithme de type PageRank, les relations entre les entités présentes dans les graphes d'entités que nous construisons à partir des textes.

8. Conclusion et perspectives

La plupart des approches pour le remplissage de formulaires en extraction d'information extraient des éléments au niveau phrastique mais exploitent assez peu le niveau discursif. Dans cet article, nous avons présenté une approche pour réaliser cette tâche conjuguant une segmentation événementielle des textes et une sélection des entités reposant sur un graphe de relations entre entités à l'échelle textuelle construit à partir de cette segmentation. Celle-ci s'appuie sur des informations temporelles pour segmenter le texte selon les événements présents en utilisant un modèle CRF afin de trouver les phrases les plus pertinentes pour remplir un formulaire donné. Ces phrases sont ensuite utilisées pour construire un graphe d'entités à partir duquel les entités relatives à l'événement d'intérêt sont sélectionnées. Nous avons proposé plusieurs stratégies pour sélectionner les entités (utilisant la position des entités, les scores de confiance des relations ou la structure du graphe, par l'utilisation de PageRank) ainsi que plusieurs façons de combiner ces stratégies (vote majoritaire ou approche hybride).

Nous avons également présenté une évaluation détaillée de notre approche sur un corpus de dépêches de presse concernant les événements sismiques. Cette évaluation a montré que notre approche a permis d'améliorer le remplissage de formulaires par rapport à une heuristique simple (mais efficace) consistant à prendre la première entité du type cherché pour remplir chaque champ du formulaire. Les résultats ont aussi montré que notre approche est particulièrement adaptée pour les documents mentionnant plusieurs événements de même nature. Finalement, une analyse des erreurs a montré que l'on peut encore améliorer ces résultats puisque la part d'erreurs liée à la sélection des entités reste de 21 %.

Concernant les perspectives de nos travaux, nous allons expérimenter la généralisation de notre approche de remplissage de formulaires à d'autres contextes, et plus précisément, d'autres langues et d'autres domaines. Nous avons déjà obtenu des résultats prometteurs en testant la segmentation événementielle sur un ensemble de dépêches

de presse en anglais, dans le domaine sismique, avec peu d'efforts d'adaptation nécessaires. Cette adaptation a en effet consisté à construire une table de correspondance entre chaque temps verbal de la langue source – le français – avec les temps de verbe de la langue cible – l'anglais – plusieurs temps de la langue cible pouvant faire référence à un seul temps de la langue source. Cette table étant appliquée aux phrases des documents à segmenter, les modèles développés pour le français ont pu être utilisés, mais sans exploiter les expressions temporelles. Nous ne prétendons pas néanmoins qu'une telle transposition est généralisable largement pour d'autres langues.

Pour ce qui est de la généralisation à d'autres domaines, il nous faut vérifier que les hypothèses faites sur la généricité des approches proposées – par exemple le fait que les informations temporelles sont suffisantes pour une segmentation événementielle ou le fait que la détection de relations entre entités ne dépend pas trop du type des entités – sont validées pour d'autres types d'événements. Par ailleurs, même si nous avons intégré un certain niveau de complexité en considérant des documents comportant plusieurs événements, cette prise en compte ne concerne que des événements de même type. Transposer les travaux réalisés à des documents plus hétérogènes en termes événementiels représente un niveau de difficulté supplémentaire important qui doit être également abordé dans une perspective de plus long terme.

Dans le prolongement de travaux déjà effectués (Jean-Louis *et al.*, 2011b), nous souhaiterions également étudier dans quelle mesure une approche de type *supervision distante* pourrait être applicable à l'étape d'extraction de relations intraphrastiques afin de limiter le recours à des exemples annotés. Nous planifions de réaliser des expérimentations relatives à la généralisation à de nouveaux domaines en utilisant les corpus des campagnes d'évaluation MUC. Enfin, l'insertion de cette méthode d'extraction d'information dans une application plus globale de veille doit également être approfondie et plus particulièrement, l'utilisation de la redondance d'information concernant le même événement dans plusieurs documents pour la fusion d'informations pourrait être explorée.

Remerciements

Nous remercions les analystes du Laboratoire de Détection et de Géophysique du CEA (au sein du Département Analyse, Surveillance, Environnement) pour l'annotation du corpus nous ayant permis de mener nos expérimentations.

Nous remercions également les trois relecteurs de notre article qui, par leurs remarques détaillées, nous ont permis de l'améliorer grandement ainsi que Béatrice Pelletier pour son travail attentif de relecture finale.

9. Bibliographie

Afzal N., « Complex Relations Extraction », *Conference on Language & Technology 2009 (CLT'09)*, Lahore, Pakistan, 2009.

Arnulphy B., Désignations nominales des événements : étude et extraction automatique dans les textes, PhD thesis, Université Paris-Sud - École Doctorale d'Informatique de Paris Sud (EDIPS) / Laboratoire LIMSI, OCT, 2012.

Besançon R., de Chalendar G., Ferret O., Gara F., Semmar N., « LIMA : A Multilingual Framework for Linguistic Analysis and Linguistic Resources Development and Evaluation », *7^{th} Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, 2010.

Besançon R., Ferret O., Jean-Louis L., « Evaluation of a Complex Information Extraction Application in Specific Domain », *8^{th} International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, 2012.

Bestgen Y., Vonk W., « Temporal Adverbials as Segmentation Markers in Discourse Comprehension », *Journal of Memory and Language*, vol. 42, n° 1, p. 74-87, 2000.

Björne J., Heimonen J., Ginter F., Airola A., Pahikkala T., Salakoski T., « Extracting Complex Biological Events with Rich Graph-Based Feature Sets », *BioNLP 2009 Workshop - Shared Task*, Boulder, Colorado, p. 10-18, 2009.

Chambers N., Jurafsky D., « Template-Based Information Extraction without the Templates », *49^{th} Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, Portland, Oregon, USA, p. 976-986, 2011.

Crowe J., « Constraint-based Event Recognition for Information Extraction », *33^{rd} Annual Meeting of the Association for Computational Linguistics (ACL'95)*, Cambridge, Massachusetts, USA, p. 296-298, 1995.

Doddington G., Mitchell A., Przybocki M., Ramshaw L., Strassel S., Weischedel R., « The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation », *4^{th} Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, p. 837-840, 2004.

Eberle K., « On Representing the Temporal Structure of a Natural Language Text », *14^{th} International Conference on Computational Linguistics (COLING'02)*, Nantes, France, p. 288-294, 1992.

Feng D., Burns G., Hovy E., « Extracting Data Records from Unstructured Biomedical Full Text », *EMNLP-CoNLL'07*, Prague, Czech Republic, p. 837-846, 2007.

Goertzel B., Pinto H., Heljakka A., Ross M., Pennachin C., Goertzel I., « Using Dependency Parsing and Probabilistic Inference to Extract Relationships between Genes, Proteins and Malignancies Implicit Among Multiple Biomedical Research Abstracts », *HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, New York, USA, p. 104-111, 2006.

Grishman R., Sundheim B., « Message Understanding Conference-6 : A Brief History », *16^{th} International Conference on Computational linguistics (COLING'96)*, Copenhagen, Denmark, p. 466-471, 1996.

Gu Z., Cercone N., « Segment-based hidden Markov models for information extraction », *21^{st} International Conference on Computational Linguistics and 44^{th} Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, p. 481-488, 2006.

Hirohata K., Okazaki N., Ananiadou S., Ishizuka M., « Identifying Sections in Scientific Abstracts using Conditional Random Fields », *Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*, Hyderabad, India, p. 381-388, 2008.

Ho-Dac L.-M., Péry-Woodley M.-P., « Temporal adverbials and discourse segmentation revisited », *7^{th} International Workshop on Multidisciplinary Approaches to Discourse 2008*

(MAD 08) - Linearisation and Segmentation in Discourse, Lysebu, Oslo, Norway, p. 65-77, 2008.

Jean-Louis L., Besançon R., Ferret O., « Text Segmentation and Graph-based Method for Template Filling in Information Extraction », *5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, Chiang Mai, Thailand, p. 723-731, 2011a.

Jean-Louis L., Besançon R., Ferret O., « Using Temporal Cues for Segmenting Texts into Events », *7th International Conference on Natural Language Processing (IceTAL 2010)*, Springer Berlin / Heidelberg, p. 150-161, 2010.

Jean-Louis L., Besançon R., Ferret O., Durand A., « A Weakly Supervised Approach for Large-scale Relation Extraction », *3rd International Conference on Knowledge Discovery and Information Retrieval (KDIR 2011)*, Paris, France, p. 94-103, 2011b.

Ji H., Grishman R., Trang Dang H., « Overview of the TAC 2010 Knowledge Base Population Track », *Third Text Analysis Conference (TAC 2010)*, Gaithersburg, Maryland, USA, 2010.

Kitani T., Eriguchi Y., Hara M., « Pattern Matching and Discourse Processing in Information Extraction from Japanese Text », *Journal of Artificial Intelligence Research*, vol. 2, p. 89-110, 1994.

Lafferty J. D., McCallum A., Pereira F. C. N., « Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data », *Eighteenth International Conference on Machine Learning (ICML'01)*, San Francisco, CA, USA, p. 282-289, 2001.

Laporte E., Nakamura T., Voyatzi S., « A French Corpus Annotated for Multiword Expressions with Adverbial Function », *6th Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Maroc, p. 48-51, 2008.

Liu Y., Shi Z., Sarkar A., « Exploiting Rich Syntactic Information for Relationship Extraction from Biomedical Articles », *NAACL-HLT'07, short paper session*, Rochester, New York, p. 97-100, 2007.

Mansuri I. R., Sarawagi S., « Integrating Unstructured Data into Relational Databases », *22nd International Conference on Data Engineering (ICDE'06)*, Washington, USA, p. 29-40, 2006.

McDonald R., Pereira F., Kulick S., Winters S., Jin Y., White P., « Simple algorithms for complex relation extraction with applications to biomedical IE », *ACL 2005*, Ann Arbor, Michigan, USA, p. 491-498, 2005.

Mihalcea R., « Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization », *42st Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, Barcelona, Spain, 2004.

Miwa M., Sætre R., Kim J.-D., Tsujii J., « Event Extraction with Complex Event Classification Using Rich Features », *Journal of Bioinformatics and Computational Biology*, vol. 8, n^{o} 1, p. 131-146, 2010.

Naughton M., « Exploiting Structure for Event Discovery Using the MDI Algorithm », *45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, Prague, Czech Republic, p. 31-36, 2007.

Patwardhan S., Riloff E., « Effective Information Extraction with Semantic Affinity Patterns and Relevant Regions », *EMNLP-CoNLL'07*, Prague, Czech Republic, p. 717-727, 2007.

Popescu A., Grefenstette G., Moëllic P. A., « Gazetiki : automatic creation of a geographical gazetteer », *8th ACM/IEEE-CS Joint conference on Digital Libraries (JCDL '08)*, p. 85-93, 2008.

Pustejovsky J., Knippen R., Littman J., Sauri R., « Temporal and Event Information in Natural Language Text », *Language Resources and Evaluation*, vol. 39, n° 2-3, p. 123-164, 2005.

Rabiner L., « A tutorial on Hidden Markov Models and selected applications in speech recognition », *Proceeding of the IEEE*, vol. 77, n° 2, p. 267-290, 1989.

Reichart R., Barzilay R., « Multi-Event Extraction Guided by Global Constraints », *2012 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (HLT-NAACL 2012)*, Montréal, Canada, p. 70-79, 2012.

Riedel S., McCallum A., « Fast and Robust Joint Models for Biomedical Event Extraction », *2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, Edinburgh, Scotland, UK., p. 1-12, 2011.

Stevenson M., « Fact distribution in Information Extraction », *Language Resources and Evaluation*, vol. 40, n° 2, p. 183-201, 2006.

Turmo J., Ageno A., Català N., « Adaptive information extraction », *ACM Computer Surveys*, vol. 38, n° 2, p. 1-47, 2006.

Warnier P., Nédellec C., « Sentence Filtering for BioNLP : Searching for Renaming Acts », *BioNLP 2011 Workshop - Shared Task*, Portland, Oregon, USA, p. 121-129, 2011.

Wick M., Culotta A., McCallum A., « Learning Field Compatibilities to Extract Database Records from Unstructured Text », *EMNLP'06*, Sydney, Australia, p. 603-611, 2006.

Yamron J. P., Carp I., Gillick L., Lowe S., van Mulbregt P., « A Hidden Markov Model Approach to Text Segmentation and Event Tracking », *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'98)*, Seattle, p. 333-336, 1998.