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
F-91191 Gif-sur-Yvette, France
{ludovic.jean-louis,romaric.besancon,olivier.ferret}@cea.fr

RÉSUMÉ

Dans les systèmes d'extraction d'information sur des événements, une tâche importante est le remplissage automatique de formulaires regroupant les informations sur un événement donné à partir d'un texte non structuré. Ce remplissage de formulaire peut s'avérer difficile lorsque l'information est dispersée dans tout le texte et mélangée à des éléments d'information liés à un autre événement similaire. Nous proposons dans cet article une approche en deux étapes pour ce problème : d'abord une segmentation du texte en événements pour sélectionner les phrases relatives au même événement ; puis une méthode de sélection dans les phrases sélectionnées des entités liées à l'événement. Une évaluation de cette approche sur un corpus annoté de dépêches dans le domaine des événements sismiques montre un F-score de 72% pour la tâche de remplissage de formulaires.

ABSTRACT

A Graph-Based Method for Template Filling in Information Extraction

In event-based Information Extraction systems, a major task is the automated filling from unstructured texts 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 a different event. 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. Using an evaluation of this model based on an annotated corpus for earthquake events, we achieve a 72% F-measure for the template-filling task.

MOTS-CLÉS : Extraction d'information, segmentation de texte, remplissage de formulaires.

KEYWORDS: Information Extraction, Text Segmentation, Template Filling.

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

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 29–42,
Grenoble, 4 au 8 juin 2012.

templates) concernant des événements. Ces formulaires permettent de mettre en évidence une information spécifique à un type d'événement considéré et d'ignorer tout autre type d'information non pertinente. La figure 1 donne un exemple du remplissage d'un formulaire à partir du texte d'une dépêche de presse.

Texte	***Templates***
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

FIG. 1 – Exemple de remplissage de formulaire

Les problèmes soulevés par la réalisation d'un système d'extraction d'information pour le remplissage de formulaire comptent en particulier l'identification des entités nommées ou 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 formulaire. 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 de la figure 1) en s'appuyant sur les mentions d'événements ; des heuristiques relatives au type d'événement ou de texte considéré sont ensuite appliquées pour fusionner les informations extraites dans des formulaires globaux. Même si ce type d'approche est largement utilisé, elle 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 de texte relatives à un événement ou un type d'événements 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. Les notions de temps et d'événement étant fortement liées, cette segmentation s'appuie

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.

2 Motivation et état de l'art

Le remplissage de formulaire 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 (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 inter-phrastique. Cette idée est nommée « hypothèse de la phrase seule » (*single sentence assumption*) par (Stevenson, 2006), qui rapporte que seulement 60% des faits mentionnés dans les corpus MUC (MUC 4-6-7) 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 nécessitent l'usage de techniques d'inférences inter-phrastiques pour les extraire.

Peu d'approches ont été proposées pour faire de l'extraction d'information à un niveau discursif sans lien étroit avec le domaine abordé. Parmi elles, (Gu et Cercone, 2006) et (Patwardhan et Riloff, 2007) sont les plus proches de l'approche présentée ici. (Gu et Cercone, 2006) définit une approche à base de modèles de Markov cachés, d'une part pour identifier les unités de textes (phrases) pertinentes pour le remplissage de formulaire, et d'autre part pour faire l'extraction des entités dans les phrases retenues. De façon similaire, (Patwardhan et Riloff, 2007) propose 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 des premières approches pour l'extraction de relations *n*-aires vient du domaine biomédical (McDonald *et al.*, 2005) et a ensuite été appliquée dans le domaine des mouvements de personnel dans les entreprises (Afzal, 2009). D'autres travaux s'attaquent 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 amène à prendre en compte des relations inter-phrastiques entre entités (Wick *et al.*, 2006; Mansuri et Sarawagi, 2006; Feng *et al.*, 2007).

3 Description de l'approche

Le cadre applicatif de la méthode d'extraction d'événements présentée dans cet article se situe dans un contexte de veille, dans lequel les utilisateurs ne sont en général intéressés que par les événements les plus récents. Dans ce contexte, notre but est de synthétiser, à partir de dépêches de presse, les informations relatives aux événements récents dans un tableau de bord. Néanmoins, les articles font en général référence à plusieurs événements comparables, en général 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. Nous utilisons une stratégie en deux étapes pour extraire cette information (Jean-Louis *et al.*, 2011) :

– une segmentation du texte en événements : les informations relatives aux événements peuvent se trouver sur plusieurs phrases. Par conséquent, nous devons découper le texte en segments homogènes du point de vue événementiel. Ces segments regroupent fréquemment des phrases non-contiguës car la structure des articles fait souvent des aller-retours entre l'événement principal et un ou plusieurs événements passés ;
– le remplissage des formulaires : puisque les segments événementiels couvrent plus d'une phrase, la probabilité d'y trouver des relations complexes (impliquant un grand nombre d'entités) est plus forte que dans une seule phrase. Nous devons donc trouver dans ces segments quelles entités sont susceptibles d'être impliquées dans des relations complexes.

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 : de façon assez liée à un domaine particulier dans des travaux comme (Gu et Cercone, 2006; Patwardhan et Riloff, 2007), avec des méthodes reposant sur des modèles très lexicalisés ; à l'inverse, en ne s'appuyant que sur la logique d'enchaînement des types d'événements dans (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.

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** : référence à l'événement principal du texte ; **Événement secondaire** : référence à un événement secondaire du texte, sans distinction de l'événement particulier s'il en existe plusieurs ; **Contexte** : sans référence à un événement.

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 les indices repérables au niveau phrastique mais également par les catégories et les indices des phrases précédentes. Notre approche se focalise ainsi sur la capture des relations entre les changements événementiels et les changements de cadre temporel, manifestées par exemple par le passage du passé composé vers plus-que-parfait accompagnant la transition de l'événement principal à un événement secondaire dans le texte de la figure 2.

Pour la classification des phrases, nous avons donc utilisé un modèle de séquences, en l'occurrence de type CRF linéaire (Champs Conditionnels Aléatoires (Lafferty *et al.*, 2001)), s'appuyant sur les traits de nature temporelle suivants. *Temps des verbes* : un trait binaire est associé à chaque temps possible et activé dès que la phrase contient au moins un verbe du temps correspondant ; *date* : la présence d'une date est souvent le signe de la présence d'un événement différent de celui de la phrase précédente ; *expression temporelle* : ce trait marque la présence d'une expression temporelle, telle que *ces dernières années, au début de l'année*, souvent associée au caractère général d'un propos. Par ailleurs, les dépendances de succession entre les différents statuts événementiels sont prises en compte par le caractère linéaire de notre modèle CRF. Ce modèle est plus amplement détaillé dans (Jean-Louis *et al.*, 2010).

5 Remplissage de formulaires événementiels

Pour le remplissage des formulaires liés aux événements, nous proposons une approche à base de graphe inspirée du paradigme de l'extraction de relations complexes. Sa première étape, dite de *construction du graphe*, détecte d'abord les relations existant entre des paires de mentions d'entités ou d'événements du domaine considéré cooccurrant dans une phrase. Il construit ensuite 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. Sa seconde 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 deux étapes sont détaillées dans les deux sections suivantes.

5.1 Construction du graphe d'entités

Le graphe d'entités que nous construisons dans cette première étape caractérise à l'échelle du document la présence ou l'absence d'une relation entre chaque paire d'entités liées au type d'événements considéré (par exemple, la relation de localisation d'un événement sismique dans notre cas). Il s'agit d'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. Le poids associé à chaque arc est un score de confiance prenant ses valeurs dans l'intervalle [0,1] et évaluant le degré de certitude de la présence d'une relation entre les deux entités liées. La figure 2 donne l'exemple d'un tel graphe restreint aux entités liées à l'événement principal du document, compte tenu de notre focalisation applicative.

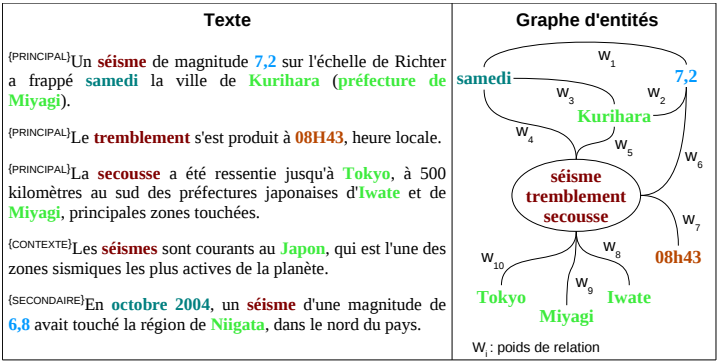

FIG. 2 – Exemple de graphe d'entités

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 détaillés dans le tableau 1.

– *LEXI-BASE* : même ensemble de traits lexicalisés que (Afzal, 2009) ;
– *LEXI-SYN* : ensemble de traits conjuguant traits lexicalisés et traits syntaxiques, dans le prolongement de (Liu *et al.*, 2007)[1] ;
– *NON-LEXI-SYN* : même ensemble de traits que *LEXI-SYN*, à l'exception des traits lexicalisés.

Comme dans (McDonald *et al.*, 2005) et (Liu *et al.*, 2007), le poids associé à chaque relation trouvée est le score de confiance du classifieur l'ayant mise en évidence, ce score étant compris dans l'intervalle [0,1] pour tous les classifieurs expérimentés à la section 6.2.

La seconde 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

[1]Nous n'utilisons pas exactement les mêmes traits que (Liu *et al.*, 2007) car certains d'entre eux ne sont applicables que dans le domaine biomédical.

Traits	LEXI-BASE	LEXI-SYN	NON-LEXI-SYN
Type d'entité de E1 ; type d'entité de E2	✓	✓	✓
Catégories morpho-syntaxiques de E1 ; cat. morpho-synt. de E2	✓	✓	✓
Mots constitutifs de E1 ; mots constitutifs de E2	✓		
Bigrammes de mots de E1 ; bigrammes de mots de E2	✓	✓	
Mots situés entre E1 et E2	✓	✓	
Bigrammes de mots situés entre E1 et E2	✓	✓	
Catégories morpho-syntaxiques situées entre E1 et E2	✓	✓	✓
Nombre de mots situés entre E1 et E2	✓	✓	✓
Bigrammes de catégories morpho-syntaxiques entre E1 et E2		✓	✓
Nombre de relations syntaxiques entre E1 et E2		✓	✓
Chemin syntaxique entre E1 et E2		✓	✓
Position / un événement + catégorie morpho-syntaxique[1]		✓	✓
Nombre d'entités situées entre E1 et E2		✓	✓
Nombre de mentions d'événement entre E1 et E2		✓	✓
Catégorie morpho-syntaxique des deux mots avant/après E1		✓	✓
Catégorie morpho-syntaxique des deux mots avant/après E2		✓	✓

[1] Si E1, respectivement E2, est une mention d'événement, associe la position de E2, respectivement E1, par rapport à elle (avant ou après) et sa catégorie morpho-syntaxique.

TAB. 1 – Traits utilisés pour la classification de relations binaires

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 2). Pour les entités, la fusion se fait sur l'égalité de leur forme normalisée dans le cas des dates, heures et magnitudes et sur l'égalité de la forme trouvée dans les textes pour les lieux. 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.

5.2 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 de 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 2 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 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 formulairessur un corpus de dépêches de presse concernant les événements sismiques, corpus décrit à la section 6.1. Une évaluation différenciée de chaque étape de notre approche a été menée : les résultats de l'évaluation de la segmentation événementielle de textes sont présentés en détail dans (Jean-Louis *et al.*, 2010) et ont montré que le modèle de segmentation par CRF atteint un F-score de 92,71% pour la détection de l'événement principal, ce qui constitue une bonne base pour l'application des étapes suivantes de notre approche. Les sections 6.2 et 6.3 présentent respectivement l'évaluation de la construction du graphe d'entités et de la sélection des entités. Une évaluation plus ciblée de l'impact de la segmentation en événements sur le résultat final est présentée à la section 6.4 et une analyse des principales erreurs rencontrées et de leur répartition est présentée dans la section 6.5.

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 cadre, un formulaire est associé à un événement sismique et résume ses principales caractéristiques à savoir, la date, l'heure, le lieu, la magnitude, les coordonnées géographiques ainsi que la mention d'événement qui lui est associée (séisme, réplique, etc.)[2]. La figure 1 donne deux exemples illustratifs du formulaire que nous considérons. Notons que dans l'application visée, nous ne cherchons à extraire que les événements principaux et ne sommes donc pas intéressés par l'événement secondaire *EV2* de cette dépêche.

[2]Les dommages liés aux séismes n'ont pas été considérés car leur expression est plus variée et leur identification nécessiterait une analyse linguistique plus profonde.

[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.

FIG. 3 – Exemples positif et négatif de présence d'une relation entre deux entités

L'ensemble des expériences ont été effectuées 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 (1/3 du corpus), et pour partie de dépêches collectées sur Google Actualités (2/3 du corpus). Le corpus a été manuellement annoté par des analystes du domaine qui ont rempli manuellement les formulaires pour chaque séisme principal d'un document. Au total, les annotateurs ont identifié 2 775 entités, réparties en 6 types d'entités : mention d'événement (18%), lieux (34%), date (17%), heure (12%), magnitude (17%) et coordonnées géographiques (1%)[3].

Concernant l'analyse linguistique des documents, nous avons appliqué la chaîne de traitements linguistiques de l'analyseur LIMA (Besançon *et al.*, 2010) réalisant les étapes de tokenisation, détection des fins de phrases, désambiguïsation morpho-syntaxique, reconnaissance des temps des verbes, reconnaissance des entités nommées et analyse syntaxique.

6.2 Construction du graphe d'entités

La méthode proposée pour la construction du graphe d'entités s'appuie sur un classifieur pour déterminer la présence/absence d'une relation entre deux entités au sein d'une même phrase. Nous avons expérimenté différents types de classifieurs statistiques, utilisant chacun les trois ensembles de traits présentés à la section 5.1 (*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. Sur 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 3 illustre des relations pour les deux catégories.

Ce corpus annoté nous a servi à tester trois types de classifieurs[4] : Bayésien Naïf (*NB*), Maximum d'Entropie (*ME*) et Arbres de décision (*DT*). Nous reportons dans le tableau 2 les résultats

[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.

[4]Nous avons utilisé l'implémentation fournie par l'outil MALLET (`http://mallet.cs.umass.edu`).

obtenus par chaque algorithme, en fonction de l'ensemble de traits utilisé en termes de rappel (*R*), précision (*P*) et F1-mesure (*F*). Les résultats sont obtenus par une validation croisée (4/5 des données servent à l'entraînement et 1/5 pour le test). 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.

Ensemble de traits	**Classifieur**	**R(%)**	**P(%)**	**F(%)**
LEXI-SYN	ME	96,30	95,92	96,10
LEXI-BASE	ME	91,22	96,09	93,57
NON-LEXI-SYN	ME	91,66	94,99	93,26
LEXI-SYN	DT	89,01	96,45	92,55
LEXI-SYN	NB	93,44	90,69	92,02
NON-LEXI-SYN	DT	91,17	88,74	89,83
NON-LEXI-SYN	NB	89,58	89,23	89,37
LEXI-BASE	DT	84,35	94,70	89,16
LEXI-BASE	NB	86,73	87,86	87,27
Baseline	–	100,00	25,50	40,49

TAB. 2 – Évaluation du classifieur de relations binaires

En premier lieu, les résultats du tableau 2 montrent 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 modèles. De plus, l'ensemble de traits non lexicalisés NON-LEXI-SYN obtient des scores équivalents à ceux de l'ensemble LEXI-BASE, ce qui est intéressant pour obtenir des modèles plus génériques. Concernant les algorithmes d'apprentissage, les performances s'organisent selon la hiérarchie ME > DT > NB. 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 modèle 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.

6.3 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 3 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*).

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

Stratégie de sélection	R(%)	P(%)	F(%)
Hybride	77,55	76,87	77,15
Vote	74,93	74,27	74,54
Confiance	74,89	74,16	74,47
Position	73,40	73,06	73,17
PageRank	72,41	71,73	72,01

TAB. 3 – Évaluation du remplissage des formulaires à partir des stratégies de sélection

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 3 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.

6.4 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 (toutes les phrases font *a priori* référence au même événement si elle comporte une mention d'événement). Celle-ci est dès lors susceptible d'apporter essentiellement des perturbations dans la mesure où ses résultats ne sont nécessairement pas parfaits.

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 4 en termes de F1-mesure et agrégés pour l'ensemble des rôles du formulaire.

Concernant les documents mono-événements, les scores du tableau 4 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

	Sans segmentation		Avec segmentation	
Stratégie	S(%)	M(%)	S(%)	M(%)
Hybride	79,20	73,61	78,34	75,61
Vote	77,67	68,68	76,89	71,81
Confiance	72,55	66,07	71,79	69,10
Position	73,96	73,16	73,07	73,10
PageRank	70,92	59,72	70,67	65,32

TAB. 4 – Impact de la segmentation sur les documents mono/multi-événements (F1-mesure)

les documents multi-événements.

6.5 Analyse des erreurs

Dans la perspective d'approfondir les évaluations de nos stratégies de remplissage de formulaire, 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 d'erreurs prépondérants :

– 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 5 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*.

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

TAB. 5 – Répartition des erreurs pour le remplissage de formulaires

Le graphe montre que la stratégie de référence *Position* 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'erreur 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%*).

7 Conclusion

La plupart des approches pour l'extraction d'information s'appuient sur des éléments au niveau phrastique pour remplir automatiquement des formulaires et peu sur des informations au niveau discursif. Dans cet article, nous avons présenté une approche pour le remplissage de formulaires fondée sur une segmentation du texte et une sélection des entités s'appuyant sur un graphe global de relations entre les entités. La segmentation du texte se fait au niveau discursif et utilise 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 formulaire 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. En ce qui concerne la généralisation à d'autres domaines, nous planifions des expérimentations dans le domaine financier.

Références

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

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

Chambers, N. et Jurafsky, D. (2011). Template-Based Information Extraction without the Templates. *In 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, pages 976–986, Portland, Oregon, USA.

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

Feng, D., Burns, G. et Hovy, E. (2007). Extracting Data Records from Unstructured Biomedical Full Text. *In EMNLP-CoNLL'07*, pages 837–846, Prague, Czech Republic.

GOERTZEL, B., PINTO, H., HELJAKKA, A., ROSS, M., PENNACHIN, C. et GOERTZEL, I. (2006). Using Dependency Parsing and Probabilistic Inference to Extract Relationships between Genes, Proteins and Malignancies Implicit Among Multiple Biomedical Research Abstracts. *In HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology*, pages 104–111, New York, USA.

GRISHMAN, R. et SUNDHEIM, B. (1996). Message Understanding Conference-6 : A Brief History. *In 16th International Conference on Computational linguistics (COLING'96)*, pages 466–471, Copenhagen, Denmark.

GU, Z. et CERCONE, N. (2006). Segment-based hidden Markov models for information extraction. *In 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 481–488, Sydney, Australia.

JEAN-LOUIS, L., BESANÇON, R. et FERRET, O. (2010). Using temporal cues for segmenting texts into events. *In 7th International Conference on Natural Language Processing (IceTAL 2010)*, pages 150–161. Springer Berlin / Heidelberg.

JEAN-LOUIS, L., BESANÇON, R. et FERRET, O. (2011). Text segmentation and graph-based method for template filling in information extraction. *In 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 723–731, Chiang Mai, Thailand.

JI, H., GRISHMAN, R. et TRANG DANG, H. (2010). Overview of the TAC 2010 Knowledge Base Population Track. *In Third Text Analysis Conference (TAC 2010)*, Gaithersburg, Maryland, USA.

LAFFERTY, J. D., MCCALLUM, A. et PEREIRA, F. C. N. (2001). Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data. *In Eighteenth International Conference on Machine Learning (ICML'01)*, pages 282–289, San Francisco, CA, USA.

LIU, Y., SHI, Z. et SARKAR, A. (2007). Exploiting Rich Syntactic Information for Relationship Extraction from Biomedical Articles. *In NAACL-HLT'07, short paper session*, pages 97–100, Rochester, New York.

MANSURI, I. R. et SARAWAGI, S. (2006). Integrating unstructured data into relational databases. *In 22nd International Conference on Data Engineering (ICDE'06)*, pages 29–40, Washington, USA.

MCDONALD, R., PEREIRA, F., KULICK, S., WINTERS, S., JIN, Y. et WHITE, P. (2005). Simple algorithms for complex relation extraction with applications to biomedical IE. *In ACL 2005*, pages 491–498, Ann Arbor, Michigan, USA.

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

NAUGHTON, M. (2007). Exploiting Structure for Event Discovery Using the MDI Algorithm. *In 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pages 31–36, Prague, Czech Republic.

PATWARDHAN, S. et RILOFF, E. (2007). Effective Information Extraction with Semantic Affinity Patterns and Relevant Regions. *In EMNLP-CoNLL'07*, pages 717–727, Prague, Czech Republic.

STEVENSON, M. (2006). Fact distribution in Information Extraction. *Language Resources and Evaluation*, 40(2):183–201.

TURMO, J., AGENO, A. et CATALÀ, N. (2006). Adaptive information extraction. *ACM Computer Surveys*, 38(2):1–47.

WICK, M., CULOTTA, A. et MCCALLUM, A. (2006). Learning Field Compatibilities to Extract Database Records from Unstructured Text. *In EMNLP'06*, pages 603–611, Sydney, Australia.