TALN 2010, Montréal, 19–23 juillet 2010

Utilisation d'indices temporels pour la segmentation événementielle de textes

Ludovic Jean-Louis Romaric Besançon Olivier Ferret
CEA, LIST, Laboratoire Vision et Ingénierie des Contenus
Fontenay-aux-Roses, F-92265, France.
{ludovic.jean-louis,romaric.besancon,olivier.ferret}@cea.fr

Résumé. Dans le domaine de l'Extraction d'Information, une place importante est faite à l'extraction d'événements dans des dépêches d'actualité, particulièrement justifiée dans le contexte d'applications de veille. Or il est fréquent qu'une dépêche d'actualité évoque plusieurs événements de même nature pour les comparer. Nous proposons dans cet article d'étudier des méthodes pour segmenter les textes en séparant les événements, dans le but de faciliter le rattachement des informations pertinentes à l'événement principal. L'idée est d'utiliser des modèles d'apprentissage statistique exploitant les marqueurs temporels présents dans les textes pour faire cette segmentation. Nous présentons plus précisément deux modèles (HMM et CRF) entraînés pour cette tâche et, en faisant une évaluation de ces modèles sur un corpus de dépêches traitant d'événements sismiques, nous montrons que les méthodes proposées permettent d'obtenir des résultats au moins aussi bons que ceux d'une approche ad hoc, avec une approche beaucoup plus générique.

Abstract. One of the early application of Information Extraction, motivated by the needs for intelligence tools, is the detection of events in news articles. But this detection may be difficult when news articles mention several occurrences of events of the same kind, which is often done for comparison purposes. We propose in this article new approaches to segment the text of news articles in units relative to only one event, in order to help the identification of relevant information associated to the main event of the news. We present two approaches that use statistical machine learning models (HMM and CRF) exploiting temporal information extracted from the texts as a basis for this segmentation. The evaluation of these approaches in the domain of seismic events show that with a robust and generic approach, we can achieve results at least as good as results obtained with an ad hoc approach.

Mots-clés : Extraction d'information, extraction d'événements, segmentation de textes, indices temporels, apprentissage statistique.

Keywords: Information extraction, event extraction, text segmentation, temporal cues, statistical machine learning.

1 Introduction

Dès les premiers travaux en extraction d'information, tels que ceux réalisés dans le cadre des conférences MUC (*Message Understanding Conference*) (Grishman & Sundheim, 1996) et plus tard des évaluations ACE (*Automatic Content Extraction*), l'extraction des événements d'un texte s'est avérée une tâche centrale. Elle constitue en effet le point de déclenchement du processus d'extraction d'information visant à compléter des *templates* prédéfinis associés à un nombre restreint de types d'événements. Dans le contexte des attaques terroristes par exemple, un tel processus consiste à identifier le type de l'attaque (explosion, etc.), sa date, ainsi que le lieu concerné. Les informations associées aux événements dans les *templates* sont en général repérées par des entités nommées, et selon les cas, la notion d'événement se matérialise par une relation qui peut soit être directe entre entités nommées (par exemple, <Hurricane>**l'ouragan Hugo**</Hurricane> de <Date>1989</Date>), soit médiée par un verbe ou un déverbal (par exemple, <Hurricane>l'ouragan Hugo</Hurricane> qui **a dévasté** <Location>les Antilles</Location>) ou encore s'étendre au-delà d'un contexte phrastique. La plupart des travaux en extraction d'information se concentrent sur les deux premiers cas dans la mesure où l'identification d'une relation entre un événement et une entité nommée s'appuie le plus souvent sur des patrons lexico-syntaxiques ou des relations syntaxiques. Néanmoins, comme le souligne (Stevenson, 2006), une part significative de ces relations ne peuvent être identifiées qu'à un niveau discursif. Jusqu'à présent, cette voie a surtout été explorée au travers de la résolution de coréférences ou de l'utilisation de connaissances sur le domaine. Dans cet article, nous abordons cette problématique par le biais de la segmentation du discours. Plus précisément, nous proposons de segmenter les textes suivant les événements qu'ils évoquent afin de diminuer les ambiguïtés de rattachement des entités.

Une telle segmentation événementielle des textes se retrouve dans quelques travaux. Elle repose souvent sur l'identification des différents constituants d'un type d'événements définis comme connaissance *a priori* sur le domaine et peut aussi intégrer des informations sur l'organisation discursive des textes. (Kitani *et al.*, 1994) distinguent ainsi deux types de structures discursives dans les articles de journaux qu'ils considèrent : l'une se présente comme une succession d'événemenents distincts tandis que l'autre s'articule autour d'un événement principal avec des références courtes à différents événements secondaires liés au premier. Cette distinction rejoint en partie celle opérée dans (Lucas, 2004), qui propose de différencier les textes à structure simple et les textes à structure complexe : les premiers sont centrés sur un seul événement décrit selon un seul point de vue ; les seconds font référence à plusieurs événements parmi lesquels se distingue un événement principal auquel se subordonnent les autres événements. (Crowe, 1995) a testé quant à lui plusieurs heuristiques de nature discursive concernant le rattachement d'une proposition à un événement dans une perspective proche de la théorie du Centrage. Enfin, (Naughton, 2007) représente une tendance un peu différente dans la mesure où l'objectif principal de la segmentation des textes y est de délimiter des segments faisant référence à des événements et des segments à caractère non événementiel. Cette segmentation est réalisée par un modèle statistique d'étiquetage des phrases en quatre types (nouvel événement, continuation d'un événement, reférence à un événement déjà mentionné, sans événement) implémenté par un automate probabiliste.

Compte tenu de l'étroite dépendance existant entre les dimensions temporelle et événementielle des textes (Pustejovsky *et al.*, 2005), l'utilisation d'indices temporels pour mettre en évidence la segmentation des textes en événements apparaît comme une voie intéressante. Les relations entre temps et segmentation du discours ont en fait été principalement abordés dans le domaine de la linguistique textuelle et de la psycho-linguistique au travers de l'étude du rôle discursif des adverbes de temps placés à l'initiale des propositions. Dans le domaine de la psycho-linguistique, (Bestgen & Vonk, 2000) montrent ainsi l'exis-

tence d'une corrélation entre la présence de tels adverbes et des changements de thème tandis que du point de vue linguistique, (Ho-Dac & Péry-Woodley, 2008) décrivent une situation plus complexe dans laquelle le rôle discursif de ces adverbes est dépendant du type des textes.

Dans cet article, nous nous focaliserons sur l'utilisation d'indices temporels pour la segmentation événementielle en nous intéressant particulièrement au cas de textes à structure complexe. Nous présenterons d'abord les principes généraux de l'approche que nous utilisons pour l'extraction d'événements en section 2. Dans la section 3, nous détaillerons les techniques utilisées pour la segmentation de textes en événements avant de rendre compte de l'évaluation de ces techniques en section 4 pour des événements dans le domaine sismique.

2 Principes et objectifs

L'extraction d'événements que nous opérons prend place dans un contexte de veille au sein duquel les utilisateurs sont essentiellement intéressés par les événements les plus récents : le but est alors de fournir à ces utilisateurs une information synthétique relative à ces événements récents à partir de leur évocation dans des dépêches d'actualité[1]. Or, une des caractéristiques des dépêches de presse est de faire fréquemment référence à plusieurs événements de même nature, en général pour donner des points de comparaison par rapport à l'événement faisant l'objet de l'actualité. Dans notre cas, les autres événements rapportés dans les dépêches ne nous intéressent pas en tant que tels et sont vus comme des sources de perturbation de l'extraction des informations relatives à l'événement principal.

La démarche que nous proposons pour aborder ce problème distingue deux étapes :
– segmenter le texte en fonction des différents événements qui y sont évoqués. Comme il est fréquent dans les structures des dépêches d'actualité de faire des allers-retours entre des événements passés et présents, les segments ne sont pas forcément contigus ;
– rattacher à l'événement d'un segment les entités nommées de ce segment qui lui sont associées. Dans notre cas, ce rattachement est opéré seulement pour les segments liés à l'événement principal.

Cette démarche est illustrée au niveau de la figure 1 sur un exemple de texte concernant un événement sismique. L'étude que nous présentons ici se focalise principalement sur la segmentation des textes en

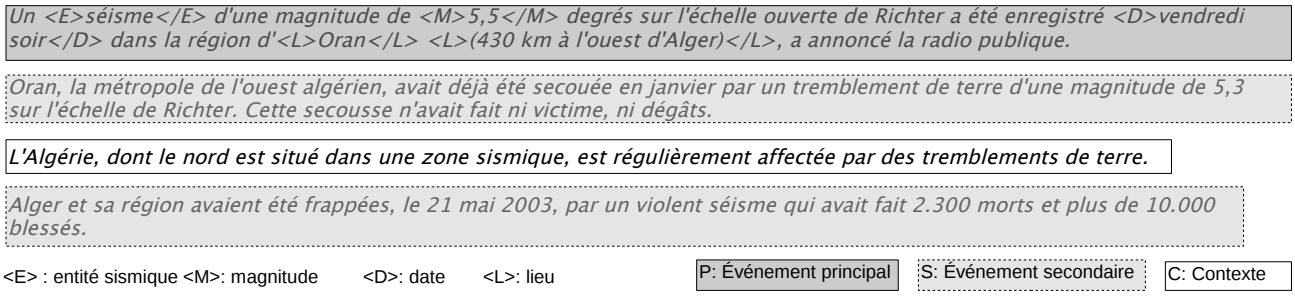

FIG. 1 – Annotation des textes en entités nommées et en événements

événements. Dans ce cadre, un texte est vu comme une séquence de phrases où chaque phrase est associée

[1]Nous nous différencions en cela des approches classiques d'extraction générale d'événements où l'on cherche à associer des informations à tous les événements évoqués dans un texte.

à un événement ou à une absence d'événement[2]. Comme nous nous intéressons au repérage des entités associées au seul événement principal de la dépêche, nous ne distinguons pas les autres événements. Ainsi, nous proposons de classer les phrases selon les trois catégories suivantes :
– **Événement principal** : toutes les phrases faisant référence à l'événement principal du texte ;
– **Événement secondaire** : toutes les phrases en rapport avec une information associée à un événement différent de l'événement principal ;
– **Contexte** : toutes les phrases n'appartenant ni à l'*Événement principal* ni à un *Événement secondaire*.
Pour réaliser cette classification, nous supposons que les critères les plus intéressants reposent non sur la seule nature des phrases mais sur leur enchaînement dans le cadre d'une structure textuelle. Au niveau linguistique, le passage d'une catégorie d'événement à une autre peut s'observer de plusieurs façons : l'usage d'un marqueur temporel différent de celui de l'événement principal (nouvelle date par exemple) ou l'emploi d'un temps grammatical différent (antérieur) à celui de l'événement principal. Ainsi, dans l'exemple donné à la figure 1, les successions de temps *passé composé/plus-que-parfait/présent/plus-que-parfait* correspondent à des changements de contexte événementiel *principal/secondaire/contexte/secondaire*. Ces changements sont de plus renforcés par une mention explicite de la date de ces événements dans les deuxième et dernière phrases, qui correspondent à un événement secondaire, et dans la troisième phrase par l'expression temporelle de périodicité *régulièrement*. Nous souhaitons capturer la dépendance entre ces changements de cadre temporel et d'événement par des techniques d'apprentissage automatique afin d'effectuer une segmentation événementielle.

3 Méthodes de segmentation en événements

Nous avons choisi de traiter la problématique de la segmentation de textes en événements comme un problème de classification de séquences, dans lequel l'objectif est d'attribuer un type d'événement à chaque phrase ou proposition. Un modèle graphique d'annotation de séquences paraît donc particulièrement adapté : nous décrivons ici deux approches de segmentation en événements fondées respectivement sur les modèles de Markov Cachés (*Hidden Markov Model*, ou HMM) et les Champs Aléatoires Conditionnels (*Conditional Random Fields*, ou CRF).

3.1 Pré-traitement 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, reconnaissance des temps de verbes, reconnaissance des entités nommées. Cette chaîne de traitement est mis en œuvre par l'analyseur linguistique LIMA présenté dans (Besançon *et al.*, 2010 à paraître).

3.2 Modèle HMM

Le modèle HMM est un modèle de classification de séquences (Rabiner, 1989) très largement utilisé en TAL (reconnaissance d'entités nommées, désambiguïsation morpho-syntaxique) et qui a déjà été appliqué

[2]L'hypothèse « une phrase = un événement » est bien entendu simplificatrice mais ne s'avère en pratique pas trop réductrice dans les domaines d'applications que nous avons testés.

pour segmenter de textes, par exemple pour 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 qui est associée à un texte donné, considéré comme séquence de phrases.

Nous faisons l'hypothèse que la segmentation est un processus markovien, c'est-à-dire que l'état associé à l'observable courant ne dépend que des observables précédents et de l'état précédent : nous proposons d'utiliser les marqueurs temporels (temps grammaticaux) comme observables, les catégories d'événements constituant les états cachés. Les matrices de transitions (une pour les états, une pour les observables) sont obtenues à partir d'un corpus de textes annotés manuellement. Une illustration du modèle HMM que nous utilisons est proposée dans la figure 2.

FIG. 2 – Illustration de la segmentation de textes en événements avec le modèle HMM

Une contrainte due à l'utilisation des HMM est que pour une séquence d'observation donnée, le calcul de la séquence d'état correspondant ne considère que l'état précédent et ne prend pas en compte les dépendances qui existent entre l'état précédent et la séquence d'observation et ne permet pas d'intégrer des critères plus variés. Pour y remédier, nous avons essayé un modèle utilisant les CRF, décrit dans la section suivante.

3.3 Modèle CRF

Depuis leur introduction en 2001, les CRF (Lafferty *et al.*, 2001) ont été 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 fait que 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 (on calcule $P(y|x)$) 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. Une définition plus formelle des CRF est donnée par (en reprenant (Hirohata *et al.*, 2008)) :

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

$$F(y,x) = \sum_i f(y,x,i) \quad (2)$$

$$Z_\lambda(x) = \sum_y exp(\lambda.F(y,x)) \quad (3)$$

où $F(y, x)$ est un vecteur ayant pour composantes les valeurs des traits pour chaque séquence i de la séquence d'entrée, 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 proposons d'intégrer les traits suivants à notre modèle de segmentation en événements :
- **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*.

4 Évaluation

Cette section présente les résultats que nous avons obtenus en appliquant les modèles HMM et CRF à la segmentation des textes en événements. Pour la mise en œuvre de ces modèles, nous avons utilisé deux implémentations de référence : NLTK[3] (HMM) et CRF++[4] (CRF). L'évaluation est proposée en deux volets : d'une part, une évaluation intrinsèque de la segmentation (les segments trouvés par la méthode sont-ils bons ?) ; d'autre part, une évaluation finale sur l'application visée, c'est-à-dire l'impact de la qualité de la segmentation sur l'extraction des informations de l'événement principal dans les dépêches.

4.1 Données

Pour l'évaluation des modèles, nous avons utilisé un corpus de 501 dépêches de presse en langue française concernant les événements sismiques. 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). Ces dépêches évoquent 142 événements sismiques principaux différents. On y retrouve à la fois des dépêches ayant une structure simple (1 seul événement) et une structure complexe (plusieurs événements) : 252 dépêches (50%) mentionnent au moins un événement secondaire. Le corpus a été manuellement annoté en entités nommées par des analystes du domaine, uniquement pour les entités liées à l'événement principal ; en revanche, les annotateurs pou-

[3] http ://www.nltk.org/
[4] http ://crfpp.sourceforge.net/

vaient annoter plusieurs entités du même type s'ils jugeaient qu'elles étaient identiquement acceptables comme information associée à l'événement (ces alternatives concernent par exemple plusieurs niveaux de granularité de l'information ; par exemple pour des noms de lieux, la ville ou le pays). Les informations associées à un événement sismique sont présentées dans le tableau 1, avec leur distribution dans le corpus. On remarque que la distribution des entités nommées n'est pas homogène : il y a beaucoup de noms de lieux (*28,64%*) et très peu de coordonnées géographiques (*0,91%*). Pour évaluer notre approche

Type d'entité	Nombre	Nature
EVENT_TYPE	499	type d'événement (séisme, tsunami ...)
LOCATION	947	lieu de l'événement
DATE	470	date de l'événement
TIME	345	heure de l'événement
MAGNITUDE	484	magnitude
DAMAGES	531	dégâts causés par l'événement
GEO_COORDINATES	30	coordonnées géographiques

TAB. 1 – Distribution des entités nommées dans le corpus de référence : 3306 entités dans 501 dépêches

de segmentation en événements, nous avons annoté une sous-partie du corpus, composée de 140 dépêches principalement sélectionnées parmi les dépêches évoquant au moins un événement secondaire. Le tableau 2 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 distincts évoqués peut monter jusqu'à 4, avec un nombre moyen de 1,66 événements secondaires évoqués par article.

1659 événements sismiques dans 140 dépêches		
Type d'événement	Nombre de phrases	Représentativité
Événement principal	1168	70%
Événement secondaire	287	17%
Contexte	213	13%

TAB. 2 – Distribution des types d'événements dans le corpus de référence

4.2 Évaluation intrinsèque de la segmentation en événements

Nous reportons dans le tableau 3 les résultats en termes de précision-rappel (notés P. et R.) pour la segmentation des textes en événements en utilisant les modèles HMM et CRF. Ces résultats sont obtenus par validation croisée en exploitant 4/5 du corpus pour la phase d'apprentissage et 1/5 pour la phase de test. Ils sont complétés par les résultats d'une heuristique ad hoc de segmentation, *HeurSeg*, issue d'une application existante d'extraction d'information dans le domaine des événements sismiques[5] et développée

[5]Cette application est actuellement utilisée par les analystes du Laboratoire de Détection et de Géophysique du CEA.

spécifiquement pour ce domaine. Cette heuristique 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 la présence d'autres entités caractéristiques du domaine entre les dates. Pour sa part, le modèle HMM exploite les seules séquences des temps des verbes contenus dans les textes. Le modèle CRF intègre tous les traits présentés à la section 3.3 ainsi que les dépendances entre la catégorie de l'événement courant et la catégorie de l'événement précédent. Il faut noter que les résultats obtenus par les modèles HMM et CRF ne sont pas directement comparables puisque les deux modèles utilisent des caractéristiques différentes pour la classification (respectivement des observations et des traits). Néanmoins, les modèles HMM et CRF obtiennent tous deux des meilleurs résultats que l'heuristique ad hoc, excepté le HMM pour les événements secondaires. On remarque que pour le HMM le seul critère utilisé n'est pas suffisant pour discriminer les événements : si l'événement principal est correctement reconnu (82,95% rappel et 93,56% de précision), les autres types d'événements le sont nettement moins. De façon globale, on peut noter que le modèle CRF permet d'obtenir une meilleure segmentation en événements, notamment de par sa capacité à intégrer un plus large ensemble d'informations.

	HeurSeg		HMM		CRF	
Type d'événement	R. (%)	P. (%)	R. (%)	P. (%)	R. (%)	P. (%)
Événement principal	82,8	64,68	82,95	93,56	98,69	87,39
Événement secondaire	23,53	43,42	37.84	9,63	52,65	95,76
Contexte	16,87	21,72	49,15	39,97	69,31	92,96

TAB. 3 – Résultats de la segmentation en événements

4.3 Évaluation de la segmentation pour l'extraction d'information

L'objectif de la segmentation en événements est de constituer des segments de texte faisant référence à un seul événement. Les segments établis sont ensuite utilisés pour rattacher les entités aux événements (le rattachement se fait à l'intérieur d'un segment). L'heuristique simple que nous appliquons pour le rattachement se fonde sur l'hypothèse que les informations contenues dans les dépêches sont organisées en fonction de leur importance dans l'actualité : les informations les plus pertinentes (généralement associées à l'événement principal) sont citées avant les informations subordonnées (associées à un événement secondaire ou au contexte). Nous utilisons donc l'heuristique suivante : *pour chaque type d'entité, on choisit la première entité trouvée dans le segment*. Pour démontrer l'intérêt de la segmentation en événements, nous reportons dans le tableau 4 les résultats du rattachement des entités à l'événement principal dans le cas où il n'y a pas de segmentation en événements (dans ce cas, on considère le document comme un seul segment) et lorsque la segmentation est réalisée par l'heuristique de segmentation *HeurSeg* de la Section 4.2 ainsi que par les modèles HMM et CRF. Il faut d'abord souligner que l'approche sans segmentation permet d'obtenir un niveau de rattachement déjà élevé (et même supérieur au HMM : +2,73% en F1-mesure) que la segmentation fondée sur l'heuristique *HeurSeg* améliore de façon significative (+6,22% en F1-mesure par rapport à l'approche basique). Moyennant les variations selon les types d'entités, le modèle à base de CRF donne pour sa part des résultats aussi bons (et même un peu meilleurs) que ceux obtenus avec la segmentation heuristique tout en offrant une approche ne dépendant pas du domaine considéré.

	Sans segmentation		HeurSeg		HMM		CRF	
Type d'entité	R. (%)	P. (%)	R. (%)	P. (%)	R. (%)	P. (%)	R. (%)	P. (%)
DAMAGES	83,51	77,88	76,29	74,37	69,85	65,14	80,15	75,3
DATE	38,41	35,87	69,31	64,99	48,93	45,6	64,38	60,12
EVENT_TYPE	82,09	81,6	79,28	78,8	59,15	58,8	76,66	76,2
GEO_COORDINATES	86,67	96,3	66,67	74,07	86,67	96,3	83,33	92,59
LOCATION	41,00	40,92	56,00	55,89	61,2	61,08	57,4	57,29
MAGNITUDE	93,54	92,96	86,25	85,89	66,67	66,25	86,67	86,13
TIME	61,05	51,22	56,4	49,24	78,78	71,5	63,37	55,47
Tous	66,58	63,5	71,02	68,63	63,44	61,15	71,65	68,82

TAB. 4 – Résultats du rattachement des entités à l'événement principal

5 Conclusion

L'objectif de cet article est l'étude de la segmentation des textes en événements, dans le but de faciliter le rattachement des entités pertinentes à l'événement principal. Dans notre approche, nous avons traité la problématique de la segmentation des textes en événements comme un problème de classification de séquences où l'objectif est de déterminer un type d'événement associé à chaque phrase. Nous avons proposé et évalué deux modèles : un modèle HMM, qui utilise comme seul critère de décision la succession des temps des verbes dans un texte et un modèle CRF, qui pour sa décision intègre en plus des indices temporels supplémentaires (expressions temporelles, dates). En évaluant les différents modèles sur un corpus de dépêches concernant les événements sismiques, nous avons montré que le modèle CRF obtient de meilleurs résultats pour la segmentation des textes en événements. De plus, nous avons vérifié l'impact de la segmentation des textes en événements sur l'identification des entités pertinentes rattachées à l'événement principal de la dépêche, et nous avons montré que le modèle à base d'apprentissage par CRF permet d'avoir des résultats équivalents (et même un peu meilleurs) à ceux obtenus avec un système utilisant une heuristique ad hoc, avec une approche a priori beaucoup plus générique.

Concernant la généricité de l'approche, nous avons fait des premiers tests encourageants sur l'application du modèle pour une autre langue (l'anglais), en utilisant directement les modèles appris du français. Concernant le domaine, nous avons utilisé un corpus de dépêches dans le domaine des événements sismiques où l'information est bien structurée. Néanmoins, nous pensons que l'approche peut donner de bons résultats dans d'autres domaines et nous pensons faire des tests dans ce sens prochainement. Enfin, une analyse plus fine des erreurs d'identification d'entités sur notre corpus d'évaluation nous a montré que la principale source d'erreur est maintenant la technique de rattachement des entités aux événements, pour laquelle une heuristique simple a été utilisée. Nous allons donc focaliser notre recherche future sur cette problématique, en exploitant à la fois des critères de densité des entités et des critères linguistiques (liens explicites de dépendance syntaxique entre les entités).

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.

Références

BESANÇON R., DE CHALENDAR G., FERRET O., GARA F. & SEMMAR N. (2010, à paraître). LIMA : A Multilingual Framework for Linguistic Analysis and Linguistic Resources Development and Evaluation. In *7th Conference on Language Resources and Evaluation (LREC 2010)*, Malta.

BESTGEN Y. & VONK W. (2000). Temporal Adverbials as Segmentation Markers in Discourse Comprehension. *Journal of Memory and Language*, **42**(1), 74–87.

CROWE J. (1995). Constraint-based Event Recognition for Information Extraction. In *33rd Annual Meeting of the Association for Computational Linguistics (ACL'95)*, p. 296–298, Cambridge, Massachusetts, USA.

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

HIROHATA K., OKAZAKI N., ANANIADOU S. & ISHIZUKA M. (2008). Identifying Sections in Scientific Abstracts using Conditional Random Fields. In *Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*, p. 381–388, Hyderabad, India.

HO-DAC L.-M. & PÉRY-WOODLEY M.-P. (2008). Temporal adverbials and discourse segmentation revisited. In *7th International Workshop on Multidisciplinary Approaches to Discourse 2008 (MAD 08) - Linearisation and Segmentation in Discourse*, p. 65–77, Lysebu, Oslo, Norway.

KITANI T., ERIGUCHI Y. & HARA M. (1994). Pattern Matching and Discourse Processing in Information Extraction from Japanese Text. *Journal of Artificial Intelligence Research*, **2**, 89–110.

LAFFERTY J. D., MCCALLUM A. & 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)*, p. 282–289, USA.

LAPORTE E., NAKAMURA T. & VOYATZI S. (2008). A French Corpus Annotated for Multiword Expressions with Adverbial Function. In *6th Conference on Language Resources and Evaluation (LREC'08)*, p. 48–51, Marrakech, Maroc.

LUCAS N. (2004). La rhétorique des dépêches de presse à travers les marques énonciatives du temps, du lieu et de la personne. In *Journée ATALA : Modéliser et décrire l'organisation discursive à l'heure du document numérique*, La Rochelle, France.

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

PUSTEJOVSKY, JAMES, KNIPPEN, ROBERT, LITTMAN, JESSICA, SAURI & ROSER (2005). Temporal and Event Information in Natural Language Text. *Computers and the Humanities*, **39**(2-3), 123–164.

RABINER L. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Readings in Speech Recognition*, p. 267–290.

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

YAMRON J. P., CARP I., GILLICK L., LOWE S. & VAN MULBREGT P. (1998). A Hidden Markov Model Approach to Text Segmentation and Event Tracking. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'98)*, p. 333–336.