Techniques d'apprentissage supervisé pour l'extraction d'événements *TimeML* en anglais et français

Béatrice Arnulphy[*], **Vincent Claveau**[†], **Xavier Tannier**[§], **Anne Vilnat**[§]

[*]*Inria - Rennes-Bretagne Atlantique,* [†]*IRISA-CNRS, Rennes, France*
beatrice.arnulphy@inria.fr vincent.claveau@irisa.fr
[§]*Univ. Paris Sud, LIMSI-CNRS, Orsay, France*
xavier.tannier@limsi.fr anne.vilnat@limsi.fr

RÉSUMÉ. *L'identification des événements au sein de textes est une tâche d'extraction d'informations importante et préalable à de nombreuses applications. Au travers des spécifications* TimeML *et des campagnes TempEval, cette tâche a reçu une attention particulière ces dernières années, mais aucun résultat de référence n'est disponible pour le français. Dans cet article nous tentons de répondre à ce problème en proposant plusieurs systèmes d'extraction, en faisant notamment collaborer champs aléatoires conditionnels, modèles de langues et k-plus-proches-voisins. Ces systèmes sont évalués sur le français et confrontés à l'état-de-l'art sur l'anglais. Les très bons résultats obtenus sur les deux langues valident notre approche.*

ABSTRACT. *Identifying events from texts is an information extraction task necessary for many NLP applications. Through the* TimeML *specifications and TempEval challenges, it has received some attention in the last years, yet, no reference result is available for French. In this paper, we try to fill this gap by proposing several event extraction systems, combining for instance Conditional Random Fields, language modeling and k-nearest-neighbors. These systems are evaluated on French corpora and compared with state-of-the-art methods on English. The very good results obtained on both languages validate our whole approach.*

MOTS-CLÉS : *Identification d'événements, extraction d'information, TimeML, TempEval, CRF, modèle de langues, anglais, français.*

KEYWORDS: *Event identification, information extraction, TimeML, TempEval, CRF, language modeling, English, French.*

CORIA 2014, pp. 89–104, Nancy, 18-21 mars 2014

1. Introduction

La détection d'événements dans les textes est une pierre angulaire pour de nombreuses applications d'accès à l'information (systèmes de questions-réponses, de dialogue, de fouille de textes...). Ces dernières années, cette tâche a reçu une certaine attention au travers des conférences-compétitions *TempEval*[1] (2007, 2010, 2013). Ces conférences ont organisé des challenges d'extraction d'événements en fournissant des corpus annotés au format *TimeML* (cf. sec. 2.1) dans plusieurs langues, ainsi qu'un cadre d'évaluation qui a permis d'établir des résultats de référence et permettant des comparaisons pertinentes entre différents systèmes.

Cependant, malgré le succès du challenge multilingue *TempEval-2*, aucun participant n'a proposé de systèmes pour le français, toutes tâches confondues. À ce jour, la situation est telle que :

– les quelques études dédiées à l'extraction d'événements en français ne peuvent pas être comparées entre elles ;

– les performances des systèmes développés ne peuvent être comparées à celles des systèmes état-de-l'art (développés sur l'anglais).

Le travail décrit dans cet article tente de répondre à ce double problème en proposant des systèmes pour l'identification d'événements pour le français. Ils sont évalués dans différents cadres de manière à les comparer à l'état-de-l'art, notamment les systèmes développés pour l'anglais. Plus précisément, les tâches que nous abordons sont celles de l'identification des événements et des noms marqueurs d'événements. Nos systèmes se veulent souples pour pouvoir être adaptés facilement à différentes langues ou types de données. Ils sont basés sur des techniques d'apprentissage classiques – arbres de décision, champs aléatoires conditionnels (*Conditional Random Fields*, noté CRF par la suite), k-plus-proches-voisins (noté kNN) – mais tirent parti de ressources lexicales soit existantes, soit construites de manière semi-automatique. Ces systèmes sont testés sur différents corpus d'évaluation, y compris ceux du challenge *TempEval-2*. Ils sont d'une part appliqués aux données en anglais, ce qui nous permet de positionner les performances obtenues par rapport aux meilleurs systèmes de l'état-de-l'art, et d'autre part sur le français.

L'article est structuré de la manière suivante. Dans la section 2, nous revenons tout d'abord sur ce qui constitue le cadre de notre travail, à savoir, les tâches d'extraction de *TempEval* et le schéma d'annotation TimeML. Nous présentons ensuite en section 3 une revue des systèmes de l'état-de-l'art sur ces tâches. Nous décrivons ensuite nos systèmes d'extraction (section 4) avant de présenter les résultats obtenus dans nos différentes expériences en anglais (section 5) puis en français (section 6).

1. `http://www.timeml.org/tempeval2/`

2. Extraction d'événements : le cadre TempEval

Les conférences *TempEval*, à travers les challenges qui y ont été organisés, ont offert un cadre incontournable aux travaux sur l'extraction d'événements. Ces challenges reposent résolument sur le langage de spécification temporelle *ISO-TimeML*. Dans la suite de cette section, nous présentons quelques éléments de ce langage avant de présenter plus précisément les challenges *TempEval*.

2.1. *TimeML*

La définition utilisée dans les challenges *TempEval* de ce qu'est un événement temporel suit les spécifications du langage *ISO-TimeML* (Pustejovsky *et al.*, 2010). *TimeML* a été développé pour l'annotation et la standardisation des événements et des expressions temporelles dans les langues naturelles. Selon cette norme, un événement est décrit de manière générique comme *"a cover term for situations that happen or occur"* (Pustejovsky *et al.*, 2003). Ce schéma d'annotation permet notamment de repérer dans les textes (pour le détail et des exemples, voir (Saurí *et al.*, 2005)) :

– les expressions événementielles (repérées par une balise <EVENT>), avec leur classe et attributs (temps, aspect, polarité, modalité). Il y a sept classes d'événements : aspectuel ASPECTUAL, action I_ACTION, prédicat d'état I_STATE, occurrence OCCURRENCE, perception PERCEPTION, rapport REPORTING et état STATE ;

– les expressions temporelles et leur valeur normalisée (<TIMEX3>) ;

– les relations temporelles entre événements et expressions temporelles (<TLINK>) ;

– les relations aspectuelles (<ALINK>) et modales (<SLINK>) entre événements ;

– les marqueurs linguistiques qui permettent ces relations (<SIGNAL>).

Ce schéma d'annotation a tout d'abord été utilisé pour annoter de l'anglais, puis appliqué à d'autres langues (avec quelques changements et adaptations du guide d'annotation à chaque langue traitée). Les corpus annotés en *TimeML* sont appelés corpus TimeBank : *TimeBank 1.2* (Pustejovsky *et al.*, 2006) pour l'anglais, *FR-TimeBank* (Bittar, 2010) pour le français... En pratique, il faut noter que les événements dans ces corpus sont majoritairement des verbes et des dates. Les événements de types nominaux, bien qu'importants pour de nombreuses applications, sont marginaux, ce qui est susceptible de poser des problèmes spécifiques à leur identification (cf. sections 5 et 6).

Dans cet article, nous nous concentrons sur la tâche d'identification des événements tels que définis par la balise *TimeML* <EVENT> (Verhagen *et al.*, 2010) (équivalent à la tâche B de *TempEval-2*). Un exemple d'un tel événement, tiré du corpus

annoté TimeBank-1.2, est donné ci-dessous[2] : la ligne (1) est la phrase avec deux événements annotés et les lignes (2) et (3) décrivent les attributs des événements.

(1) The financial <EVENT eid="e3" class="OCCURRENCE">assistance</EVENT> from the World Bank and the International Monetary Fund are not <EVENT eid="e4" class="OCCURRENCE">helping</EVENT>.

(2) <MAKEINSTANCE eventID="e3" eid="ei377" tense="NONE" aspect="NONE" polarity="POS" pos="NOUN"/>

(3) <MAKEINSTANCE eventID="e4" eid="ei378" tense="PRESENT" aspect="PROGRESSIVE" polarity="NEG" pos="VERB"/>

2.2. *Challenges TempEval*

Il y a eu à ce jour trois éditions des campagnes d'évaluations *TempEval* (conjointes aux conférences *SemEval*[3]).
TempEval-1[4] (Verhagen *et al.*, 2007) s'intéressait à la détection de relations entre entités préalablement annotées. Cette première édition ne proposait que des textes en anglais.
TempEval-2[5] (Verhagen *et al.*, 2010) s'est concentré sur la détection d'événements, d'expressions temporelles et de relations temporelles. Cette campagne est multilingue (comprenant notamment l'anglais, le français, l'espagnol) et les tâches ont été plus précisément définies par rapport à *TempEval-1*.
TempEval-3[6] (UzZaman *et al.*, 2013) s'inscrit dans la continuité des éditions précédentes. Il s'agit là aussi d'évaluer l'extraction d'événements et de relations temporelles mais seuls l'anglais et l'espagnol sont proposés. Par ailleurs, un des focus de ce nouveau challenge était d'évaluer l'impact de l'ajout de données d'entraînement annotées automatiquement aux données annotées manuellement.

Comme nous l'avons mentionné auparavant, dans notre étude, nous nous concentrons sur l'extraction d'événements (marqués par des verbes ou des noms) comme définis initialement dans le challenge *TempEval-2*. Par ailleurs, notre but étant de produire et d'évaluer des systèmes pour le français, nous utilisons, entre autres, les données développées dans le cadre de *TempEval-2*.

3. Travaux connexes

Différents travaux se sont intéressés à l'annotation et l'extraction automatique d'événements. La plupart l'ont fait dans un cadre spécifique à une application, avec

2. Tirée de `TimeBank-1.2/data/timeml/ABC19980108.1830.0711.html`
3. `http://semeval2.fbk.eu/semeval2.php`
4. `http://www.timeml.org/tempeval/`
5. `http://semeval2.fbk.eu/semeval2.php?location=tasks#T5`
6. `http://www.cs.york.ac.uk/semeval-2013/task1/`

une définition propre de ce qu'était un événement, notamment dans des contextes de veille (Besançon *et al.*, 2011, par exemple sur événements sismiques). Ces définitions applicatives menant souvent à des systèmes dédiés, difficiles à évaluer hors de ces cadres, ces travaux ne sont pas discutés ici. Dans cette section, nous présentons les travaux les plus proches des nôtres sur ce domaine, qu'ils aient été faits dans le cadre de *TempEval-2* ou non, mais utilisant la définition linguistiquement motivée et généraliste des événements proposée dans *TimeML*.

3.1. *Extraction d'événements définis en TimeML*

Nous mentionnons ici les travaux sur les événements au sens *TimeML* en anglais puis en français. Le système EVITA (Saurí *et al.*, 2005) a pour objectif d'extraire les événements en combinant indices linguistiques et approches statistiques, et utilise *WordNet* comme ressource externe. Ce système a été évalué sur le corpus anglais *TimeBank1.2*. Avec le système STEP, Bethard et Martin (2006) cherchent à classifier tous les items *TimeML* avec une approche par apprentissage s'appuyant sur des attributs linguistiques, sans ressources externes. Ces mêmes auteurs ont également développé deux systèmes *baseline* (MEMORIZE et une simulation d'EVITA). Bien que tous les items *TimeML* étaient recherchés, les auteurs se sont intéressés particulièrement aux événements nominaux. Ils sont arrivés à la conclusion que la détection automatique de tels événements nominaux (noms notés <event>) n'est pas triviale, du fait de la grande variété d'expressions possibles et donc du manque de données d'apprentissage.

Parent *et al.* (2008) ont travaillé à l'extraction sur le français des éléments définis par *TimeML*. Ils se sont notamment intéressés aux locutions adverbiales d'indication temporelle. Leur système repose sur un étiquetage syntaxique, sur des patrons syntaxiques définis manuellement et sur une version du lexique *VerbAction* (Tanguy et Hathout, 2002). Ils ont utilisé pour leur travail un corpus de bibliographies et de romans manuellement annotés, réalisé avant la publication du *FR-TimeBank*. À notre connaissance, il s'agit des seuls travaux sur l'identification d'événements de type *TimeML* sur le français.

3.2. *Extraction d'événements dans le cadre de TempEval-2*

Plusieurs systèmes ont été proposés dans le cadre de la compétition *TempEval-2*, travaillant pour la plupart sur l'anglais. Le système TIPSEM (Llorens *et al.*, 2010) repose sur un apprentissage artificiel par CRF et des attributs de différentes natures : lemmes, parties-du-discours, informations syntaxiques obtenues automatiquement ou par des règles écrites à la main, informations sémantiques tirées de WordNet. Ce système a obtenu les meilleurs résultats à *TempEval-2* et a été utilisé comme référence pour *TempEval-3*.

Le système EDINBURGH (Grover *et al.*, 2010) s'appuie quant à lui sur une combinaison de sous-systèmes effectuant une segmentation du texte, une reconnaissance

d'entités nommées par règles et par apprentissage artificiel, une analyse syntaxique de surface. Il exploite des informations lexicales collectées sur les données d'apprentissage et sur WordNet.
TRIOS (UzZaman et Allen, 2010) est un système basé sur de l'apprentissage artificiel par réseaux logiques de Markov exploitant les annotations de l'étiqueteur TRIPS (Allen *et al.*, 2008) (étiqueteur annotant des attributs proches de ceux *TimeML*).
Enfin, le système JU_CSE (Kumar Kolya *et al.*, 2010), le moins performant dans la compétition, consiste simplement en une collection de règles d'extraction d'événements, construites manuellement, et exploitant un étiquetage en parties-du-discours.

Tous ces systèmes et leurs performances apportent des enseignements importants. Ainsi, la plupart reposent classiquement sur des approches par apprentissage, et sans surprise, comme pour beaucoup de tâches d'extraction d'information par annotation, les CRF semblent particulièrement adaptés. Ils mettent également en exergue la nécessité de disposer d'informations sémantiques avec une couverture suffisamment importante pour gérer la variété des expressions d'événements, notamment pour les noms. Les systèmes que nous présentons dans cet article partagent donc de nombreux points communs avec ceux-ci (cf. section suivante) puisqu'ils reposent sur un apprentissage supervisé, utilisant notamment les CRF, et également des lexiques obtenus en partie automatiquement.

4. Systèmes d'extraction

Les systèmes d'extraction que nous proposons se veulent souples pour pouvoir être facilement adaptés à des nouvelles langues ou textes. Comme pour de nombreux systèmes de l'état-de-l'art, ils s'inscrivent dans un cadre standard d'apprentissage artificiel supervisé : des données annotées avec les éléments *TimeML* recherchés sont fournies pour entraîner nos systèmes, qui sont ensuite évalués sur un jeu de données de test disjoint. La tâche d'apprentissage est donc une tâche d'annotation des textes : le but des classifieurs est d'assigner une étiquette à chaque mot indiquant s'il est un événement ou non. Certains événements étant exprimés par des syntagmes composés de plusieurs mots, le schéma d'annotation BIO est utilisé (B indique le début du syntagme, I la continuité, et O sont pour les mots hors du syntagme événement). Les données d'entraînement sont des extraits de corpus annotés avec ces étiquettes pour chaque mot et sont décrits par différents attributs que nous détaillons ci-dessous, qui sont ensuite exploités par les techniques d'apprentissage que nous présentons en sous-section 4.2 et 4.3. Après la phase d'entraînement, pour annoter de nouveaux textes, à chaque mot de chaque phrase, les classifieurs inférés décident de l'étiquette la plus probable en fonction des attributs du mot courant et du contexte environnant.

4.1. *Attributs*

Les attributs que nous utilisons pour décrire les mots en vue de leur étiquetage se veulent simples à obtenir automatiquement. Il s'agit tout d'abord des mots-formes,

lemmes et parties-du-dicours. Ces attributs classiques en TAL sont appelés par la suite attributs internes. Dans nos expériences, ils sont obtenus avec TreeTagger (`http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger`).

D'autres attributs, dits externes, apportent des informations lexicales qui semblent importantes pour nos tâches d'extraction. Ils sont tirés d'une part de lexiques existants, spécialisés sur les événements ou généralistes :

– pour le français, chaque mot est annoté selon qu'il appartient ou non aux lexiques *VerbAction* (Tanguy et Hathout, 2002) et *The Alternative Noun Lexicon* (Bittar, 2010). Le premier est une liste des verbes et de leurs nominalisations décrivant des actions (eg. *enfumage, réarmement*) ; le second complète cette liste en recensant les noms d'événements qui ne sont pas des déverbaux (eg. *miracle, tempête*).

– pour l'anglais, chaque mot est annoté selon qu'il appartient ou non à l'une des huit classes de synsets liées à des événements ou actions (*change, communication, competition, consumption, contact, creation, motion, stative*).

Nous utilisons également des informations lexicales construites automatiquement, à savoir les *Eventiveness Relative Weight Lexicons* (notés ERW par la suite), en suivant les principes que nous avons développés dans des travaux précédents (Arnulphy *et al.*, 2012a). Nous ne redétaillons pas les principes de construction de ces lexiques valués ici, qui peuvent être trouvés dans les références citées. Ces lexiques produisent des listes de mots associés à des valeurs indiquant leur probabilité d'exprimer un événement. Dans les expériences rapportées dans la suite de cet article, ces lexiques sont construits à partir de corpus journalistiques (corpus de dépêches AFP et Wall Street Journal).

Il faut noter que ces lexiques donnent des informations sur des mots polysémiques. Ainsi, la plupart des mots contenus dans ces lexiques peuvent dénoter une action, que l'on veut alors extraire, ou le résultat d'une action, qui n'est pas à extraire (par exemple, *enfoncement, décision*). Ils ne sont donc pas suffisants par eux seuls, mais fournissent des indices intéressants à exploiter dans des méthodes plus complexes.

4.2. *Systèmes par CRF et par arbres de décision*

Comme nous l'avons expliqué, la tâche d'extraction des événements est vue comme une tâche d'annotation pour laquelle nous inférons un classifieur à partir de données d'entraînement annotées. Parmi les nombreuses techniques d'apprentissage possibles pour ce faire, nous nous sommes intéressés dans un premier temps à deux d'entre elles, classiquement utilisés pour ce type de tâche : les champs aléatoires conditionnels (CRF), utilisées par exemple par Llorens *et al.* (2010), et les arbres de décisions (DT), qui ont montré de bonnes performances dans nos précédents travaux (Arnulphy, 2012).

Les DT que nous utilisons pour nos expériences sont ceux inférés par C4.5 (Quinlan, 1993) tels qu'implémentés dans WEKA (Hall *et al.*, 2009). L'intérêt des DT est leur capacité à manipuler des attributs de différente nature : nominaux

(parties-du-discours), booléens (appartenance aux lexiques), numériques (valeur dans les lexiques ERW)... Pour prendre en compte l'aspect séquentiel des textes, chaque mot est décrit par ses attributs (cf. sec. 4.1) et ceux des deux mots précédents et suivants (relativement à la quantité de données d'apprentissage, l'utilisation de contextes plus longs introduit des attributs trop variables pour être informatifs).

Les CRF (Lafferty *et al.*, 2001) sont désormais des outils incontournables des tâches d'annotation. À l'inverse des DT, ces modèles graphiques prennent naturellement en compte les dépendances séquentielles de nos données textuelles. En revanche, dans la plupart des implémentations, il n'est pas possible d'utiliser des attributs numériques. L'échelle des valeurs ERW est donc découpée en 10 portions de largeurs égales, et transformés en attributs nominaux à 10 valeurs. Dans nos expériences, nous utilisons l'implémentation WAPITI (Lavergne *et al.*, 2010) qui a montré sa robustesse sur de nombreuses tâches de TAL.

4.3. *Le système combiné CRF-kNN*

Les deux systèmes précédents sont des approches relativement classiques de l'extraction d'information. Nous proposons un système supplémentaire, s'appuyant également sur les CRF, mais qui cherche à en combler certaines limites. En effet, les CRF permettent bien de prendre en compte l'aspect séquentiel des données, mais avec un historique assez contraint. Une séquence introduisant un événement X, comme dans l'exemple 1 ci-dessous sera jugée différente de celle de l'exemple 2 du fait du décalage produit par l'insertion de "*l'événement de*". L'événement Y ne sera donc pas détectée même si la première phrase, pourtant ressemblante, est dans les données d'entraînement.

1) "*c'est à cette occasion que s'est produit X ...*"

2) "*c'est à cette occasion que s'est produit l'événement de Y ...*"

Par ailleurs, il est difficile d'intégrer des connaissances numériques (degré ERW) ou d'indiquer des synonymes possibles dans ces modèles.

Pour pallier ces problèmes, nous adjoignons aux CRF un kNN pour décider de la classe des candidats événements. Le CRF est employé comme expliqué dans la sous-section précédente, mais toutes les sorties (étiquetages) possibles et leur probabilités sont gardées. Le kNN calcule ensuite une similarité entre les instances à classer (tous les événements repérés quelles que soient leurs probabilités) et les données d'entraînement.

Dans notre cas, cette similarité est calculée avec des modèles de langues. Ces modèles permettent en effet d'attribuer une probabilité (notée P_{LM}) à une séquence de mots. Plus précisément, pour chaque candidat potentiel identifié par le CRF, sa classe C^* (événement ou non) est décidée en prenant en compte la probabilité trouvée par le CRF ($P_{CRF}(C)$), mais aussi celles issues des modèles de langue sur le candidat et sur les contextes droit (cont_D) et gauche (cont_G) du candidat. Un modèle (ensemble

de probabilités estimées sur les données) est donc appris pour chaque classe et pour chaque position (contextes droit ou gauche ou candidat) à partir des données d'entraînement. C'est-à-dire que pour chaque classe, on estime les probabilités d'occurrences des séquences de mots sur l'ensemble des contextes droit des événements de la classe considérée ; il est fait de même sur les contextes gauche et sur le événements eux-mêmes. On note ces modèles respectivement $\mathcal{M}_C$, $\mathcal{M}_C^D$ et $\mathcal{M}_C^G$. Finalement, on a :

$$C^* = \arg\max_C P_{CRF}(C) * P_{LM}(\text{cont}_G|\mathcal{M}_C^G) * P_{LM}(\text{candidat}|\mathcal{M}_C) * P_{LM}(\text{cont}_D|\mathcal{M}_C^D)$$

Dans nos expériences, les modèles de langue utilisés sont des modèles bigramme pour $\mathcal{M}_C^D$ et $\mathcal{M}_C^G$, et unigramme pour $\mathcal{M}_C$ (les candidats événements étant souvent composés d'un seul mot-forme) ; la longueur des contextes gauche et droit est de 5 mots. Ainsi, si l'exemple 2 est dans les données d'entraînement, la similarité des contextes gauche avec la séquence 1 sera suffisamment importante pour détecter l'événement.

Par ailleurs, l'intérêt des modèles de langue est aussi de pouvoir exploiter naturellement des informations lexicales au sein du lissage. En effet, pour éviter que l'absence d'un n-gramme dans les données d'entraînement produise une probabilité nulle pour toute la séquence testée, il est usuel d'assigner une probabilité, même petite, à ces n-grammes non vus. Diverses stratégies ont été proposées pour ce faire (Ney *et al.*, 1994). Pour notre part, nous utilisons une stratégie de repli (*backoff*) sur les unigrammes pour les bigrammes absents et un lissage laplacien, simple à mettre en œuvre, pour les modèles unigrammes. L'originalité de notre travail est d'utiliser aussi ce lissage pour tirer parti des informations apportées par nos lexiques. Ainsi, pour WordNet en anglais et les lexiques français, un mot absent des données d'entraînement est remplacé, le cas échéant, par un mot présent du même synset ou du même lexique. Si plusieurs mots sont éligibles pour le remplacer, celui maximisant la probabilité est choisi. Dans tous les cas, une pénalité (un facteur multiplicatif $\lambda < 1$) est appliquée à la probabilité ; elle est estimée par validation croisée sur les données d'entraînement. Pour un mot w absent des données d'entraînement d'un modèle $\mathcal{M}$, on a donc :

$$P(w|\mathcal{M}) = \lambda * max\{P(w_i|\mathcal{M}) \mid w_i, w \text{ dans le même lexique/synset }\}$$

Les valeurs ERW sont quant à elles interprétées comme des valeurs d'appartenance (les mots absents du lexique ont un score de 0). La pénalité de remplacement d'un mot par un autre de ce lexique est proportionnelle à l'écart entre les valeurs des deux mots concernés.

L'avantage de cette combinaison séquentielle est d'exploiter la capacité des CRF à isoler des syntagmes intéressants grâce à une décision multi-critères (partie-du-discours, lemmes...), et d'utiliser les modèles de langues pour un jugement plus global sur les contextes, avec une prise en compte naturelle des informations lexicales à travers les stratégies de lissage.

5. Expériences sur l'anglais

5.1. *Contexte expérimental*

Pour évaluer la performance de nos systèmes, nous reprenons les mesures utilisées dans le cadre du challenge *TempEval-2*, à savoir la précision (Pr), le rappel (Rc) et la f-mesure (F1). Ces mesures sont évaluées pour la tâche d'extraction des événements dans leur ensemble, mais aussi pour des sous-ensembles d'événements particulièrement intéressants réputés plus difficiles, à savoir les événements nominaux (exprimés par des noms ou des syntagmes dont la tête est un nom) et les événements nominaux sauf ceux de type états.

Au delà des résultats bruts, nous cherchons aussi à évaluer l'apport des différents types d'attributs ; pour cela, nous avons testé de nombreuses combinaisons. Dans les expériences rapportées nous présentons les résultats pour certaines configurations selon les informations lexicales qu'elles utilisent (internes, c'est-à-dire directement tirées des textes et/ou externes, c'est-à-dire issus de ressources comme WordNet). Plus précisément, les configurations sont les suivantes :

1) avec informations internes et sans information lexicale externe : les modèles exploitent uniquement les mots-formes, les lemmes et les parties-du-discours ;

2) avec informations internes et lexiques externes (cf. section 4.1) ;

3) cette configuration est une variante de la précédente, spécifique à l'usage de WordNet pour les événements en anglais. Les 8 classes de synsets de WordNet correspondant à des événements ou actions (cf. section 4.1) sont utilisées comme 8 attributs binaires indiquant l'absence ou la présence du mot courant dans ces synsets.

5.2. *Résultats sur l'anglais*

Parmi toutes les configurations, le tableau 1 récapitule les résultats des meilleurs modèles/configurations d'attributs. À des fins de comparaison, nous indiquons également les résultats des systèmes TIPSEM, EDINBUGH, JU_CSE, TRIOS et TRIPS obtenus lors de *TempEval-2*.

Sur ces données en anglais, les approches par CRF surpassent celles par arbres de décision, spécialement sur les événements nominaux. Par ailleurs, plusieurs tests et ajustements ont dû être effectués sur les approches par arbres de décision pour obtenir ces résultats sur les événements nominaux. En effet, le faible pourcentage de noms événements parmi l'ensemble des noms induit un très important déséquilibre de classes nécessitant de modifier les options d'élagage des arbres. Ce sont en effet 57,5 % des verbes qui sont des événements, contre seulement 7 % pour les noms. Ce problème d'équilibre de classes impacte moins les approches par CRF, ce qui explique en partie les différences de performances constatées entre méthodes. Quelle que soit la méthode, on constate également une différence de performances entre les événements nominaux (avec ou sans états) qui s'explique là encore par la faible proportion de ces

Type d'événement	Modèle	Pr	Rc	F1
tous événements	TIPSEM	0,81	**0,86**	0,83
	EDINBURGH	0,75	0,85	0,80
	JU_CSE	0,48	0,56	0,52
	TRIOS	0,80	0,74	0,77
	TRIPS	0,55	0,88	0,68
	(3) CRF-kNN	**0,86**	**0,86**	**0,86**
	(3) CRF	0,79	0,8	0,79
	(3) DT	0,73	0,71	0,72
nominaux seulement	(3) CRF-kNN	0,78	0,55	0,65
	(3) CRF	0,72	0,48	0,58
	(2) DT	0,58	0,28	0,38
nominaux sauf états	(3) CRF-kNN	0,64	0,44	0,52
	(3) CRF	0,53	0,38	0,45
	(3) DT	0,87	0,08	0,15

Tableau 1. *Performances des meilleurs modèles/configurations sur le corpus anglais TempEval-2.*

derniers. Cela impacte les données d'entraînement : les classifieurs ont relativement peu de données pour généraliser, ce qui pénalise principalement le rappel.

Les écarts de performances entre nos différentes configurations d'attributs nous apportent aussi un enseignement sur l'importance des informations lexicales, même externes, pour ces tâches d'extraction. Ce fait transparaissait déjà au travers de l'état-de-l'art (cf. section 3) mais est ici confirmé par la comparaison de systèmes complètement identiques par ailleurs.

Enfin, notre système CRF-kNN obtient les meilleurs résultats tous systèmes confondus, surpassant l'approche classique par CRF, celle par DT, et l'état-de-l'art. Ces résultats sont tous obtenus en prenant en compte les lexiques externes, ce qui souligne leur nécessité pour ces tâches. Ces bons résultats sont particulièrement intéressants puisque nos systèmes reposent uniquement sur des attributs simples à obtenir (parties-du-discours) ou issus de bases lexicales généralistes (WordNet). Ils doivent donc être facilement portables à une autre langue comme le français, qui fait l'objet de la section suivante.

6. Expériences sur le français

6.1. *Données d'évaluation et comparaison à l'anglais*

Comparativement à l'anglais, peu de corpus sont disponibles pour développer, évaluer et comparer des systèmes d'extraction d'événements en français. Parmi ceux-là, le corpus *TempEval-2* français est supposé comparable en terme de genre et d'annotation

à sa contrepartie anglaise. Tout comme le corpus anglais était tiré du *TimeBank1.2*, le corpus français est tiré du *FR-TimeBank* auquel a été ajouté des biographies. Dans des travaux précédents (Arnulphy *et al.*, 2012b), nous avons également proposé un corpus annoté pour le français. Comme le *FR-TimeBank*, il est composé de textes journalistiques. Il est donc comparable en type au corpus de *TempEval-2* mais n'est annoté que pour les désignations nominales d'événements qui ne sont pas des états (cela correspond au tag *TimeML* <EVENT class="OCCURRENCE" pos="NOUN">).

Plusieurs observations doivent être faites pour pouvoir comparer pertinemment les résultats obtenus sur ces différents corpus et avec l'anglais. Le tableau 2 présente quelques indicateurs globaux. Il en ressort que les proportions de tous les événements dans les corpus français et anglais de *TempEval-2* sont identiques : environ 2,6 par phrase. Par ailleurs, un examen détaillé montre qu'il y a plus d'événements verbaux que nominaux dans les corpus *TempEval-2*, mais relativement plus d'événements nominaux dans les deux corpus français que dans l'anglais. De plus, le corpus annoté d'Arnulphy *et al.* (2012b) est plus riche que le *TempEval-2* français en ce qui concerne les événements nominaux ; environ 90 % des événements nominaux ne sont pas des états dans le *TempEval-2* français, contre seulement 80 % en anglais.

		Nb de		
		phrases	tokens	événements
ENG	*TempEval-2*	2 382	58 299	6 186
FRA	*TempEval-2*	441	9 910	1 150
FRA	corpus de Arnulphy *et al.* (2012b)	2 414	54 110	1 863

Tableau 2. *Éléments de comparaison des corpus anglais (ENG) et français (FRA) annotés en TimeML.*

6.2. *Résultats sur le français*

Nous reprenons les mêmes configurations d'attributs 1 et 2 que pour l'anglais. Nous indiquons dans le tableau 3 les résultats des modèles et configurations les plus performantes. À des fins de comparaison, nous avons également implémenté une variante proposée dans un travail précédent (Arnulphy, 2012) pour servir de *baseline*. Cette précédente approche qui repose aussi sur les arbres de décision a donné de bons résultats, mais au contraire des approches proposées ici, celle-ci utilise des attributs plus coûteux à obtenir et moins portables, à savoir une analyse syntaxique profonde, post-éditée et exploitée à l'aide de règles construites manuellement. Cette *baseline* est notée 4. Nous rapportons également les chiffres publiés par Parent *et al.* (2008) sur leur propre corpus à titre informatif.

Dans l'ensemble, les modèles CRF obtiennent des résultats aussi bons que la technique proposée dans (Arnulphy, 2012), mais sans requérir d'informations syntaxiques et de ressources développées à la main. Sur l'extraction des événements nominaux sauf états, les résultats obtenus sont nettement meilleurs sur le corpus de (Arnulphy

Corpus	Type d'événement	Modèle	Pr	Rc	F1
TempEval-2 français	tous événements	(2) CRF-kNN	**0,87**	**0,79**	**0,83**
		(2) CRF	0,8	0,76	0,78
		(4) DT	0,78	0,77	0,78
	nominaux seulement	(2) CRF-kNN	**0,69**	0,60,	**0,64**
		(2) CRF	0,55	0,52	0,53
		(4) DT	0,58	**0,63**	0,6
	nominaux sauf états	(2) CRF-kNN	**0,65**	**0,52**	**0,58**
		(2) CRF	0,53	0,46	0,5
		(4) DT	0,57	0,49	0,53
Arnulphy *et al.* (2012b)	nominaux sauf états	(2) CRF-kNN	**0,79**	**0,63**	**0,70**
		(2) CRF	0,76	0,54	0,63
		(4) DT	0,75	0,60	0,67
Corpus Parent *et al.*	tous événements	Parent et al.	0,625	0,777	0,693
	nominaux seulement	Parent et al.	0,547	0,537	0,542

Tableau 3. *Performances des meilleurs modèles/configurations sur les corpus français TempEval-2, Arnulphy* et al. *(2012b) et Parent* et al. *(2008).*

et al., 2012b) que sur *TempEval-2* français (F1=0,63 vs. F1=0,53). Cette différence de performances met en lumière les différences entre les deux corpus français abordées précédemment. Enfin, même si la comparaison est délicate à cause de corpus différents, il est intéressant de noter que nos approches ont des performances supérieures à celles rapportées par Parent *et al.* (2008).

Les résultats obtenus pour le français sont comparables dans les grandes lignes à ceux de l'anglais. Comme pour l'anglais, l'extraction des événements nominaux est plus difficile que l'extraction de tous les types d'événements. La différence de performances entre événements nominaux et événements nominaux sauf états est néanmoins plus faible que celle constatée en anglais. Cela s'explique par les différences de proportions entre les deux langues que nous avions soulignées (cf. section 6.1). Comme pour l'anglais également, le système CRF-kNN obtient globalement les meilleurs résultats. On constate aussi que les résultats des différentes configurations d'attributs mettent de nouveau en valeur l'importance des informations lexicales.

6.3. *Influence des lexiques et des données d'entraînement*

Pour évaluer l'influence de la quantité de données d'entraînement sur les performances de notre système CRF-kNN, nous indiquons en figure 1 l'évolution de la F-mesure selon le nombre de phrases servant à l'entraînement. À des fins de comparaison, nous indiquons les performances du système CRF simple pour mesurer l'apport des modèles de langue du kNN. Nous testons deux configurations, avec informations lexicales et sans informations lexicales externes.

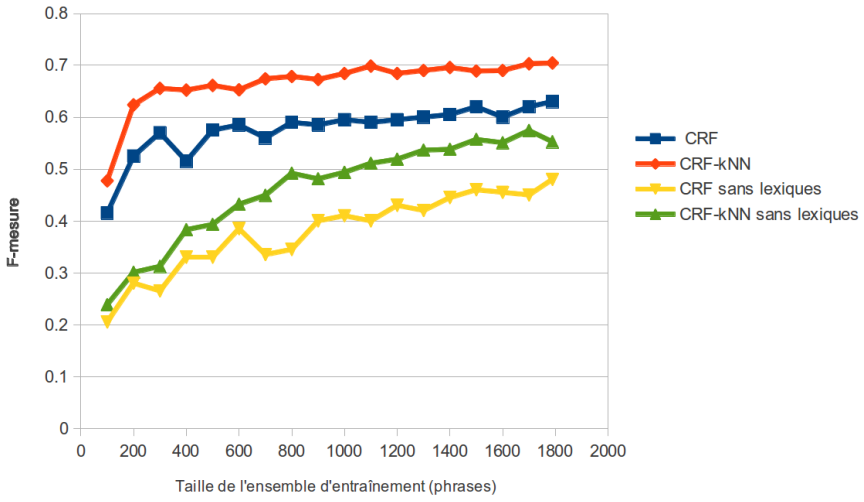

Figure 1. *Performances (F-mesure) des modèles CRF selon le nombre de phrases utilisées en entraînement.*

Plusieurs observations en ressortent. D'une part, l'apport de la combinaison des CRF avec les kNN est net, quel que soit le nombre de phrases d'entraînement. Les modèles de langues améliorent également les performances des CRF que ce soit avec ou sans lexiques. Sans informations lexicales externes, la progression de la F-mesure est bien sûr constante selon le nombre de phrases d'entraînement. En revanche, l'utilisation des ressources lexicales atténue cet effet, la progression est importante pour de petits ensembles d'entraînement mais quasiment linéaire ensuite. Cela signifie qu'un petit ensemble d'entraînement suffit dès lors que des ressources lexicales sont utilisées.

7. Conclusion

Dans cet article, nous nous sommes intéressés à l'extraction d'événements tels que définis par le standard *TimeML* et mis en oeuvre dans les challenges *TempEval*. L'absence de participation sur le corpus français de *TempEval*, ainsi que le fait que les études existantes aient été faits sur des corpus différents ou des tâches légèrement différentes, font qu'il était difficile d'établir l'état-de-l'art pour le français. Nous avons tenté de répondre à ces problèmes en proposant plusieurs systèmes que nous avons évalués sur des corpus en français, mais aussi sur des corpus en anglais pour valider leur bien-fondé en les comparant aux approches existantes qui sont, au contraire du français, bien répertoriées pour l'anglais.

Les trois systèmes que nous avons proposés sont classiquement fondés sur des techniques d'apprentissage artificiel. Parmi ceux-ci, la combinaison des CRF avec un kNN basé sur des modèles de langue obtient les meilleurs résultats et surpassent les meilleurs systèmes de l'état-de-l'art. Cette combinaison tire en effet le meilleur parti des deux techniques combinées et permet également de prendre en compte les informations lexicales externes assez simplement. Globalement, les bons résultats obtenus sur l'anglais positionnent ceux sur le français comme pouvant servir de *baseline* raisonnable pour de futurs travaux. Parmi ces perspectives, l'extraction des autres types d'indicateurs temporelles et des relations entre événements est à l'étude. Il sera intéressant d'étudier l'adaptation de notre méthode CRF-kNN à ces tâches et à d'autres tâches d'extraction d'informations.

8. Bibliographie

Allen J. F., Swift M., de Beaumont W., « Deep semantic analysis of text », *Proceedings of the 2008 Conference on Semantics in Text Processing*, STEP '08, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 343-354, 2008.

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, 2012.

Arnulphy B., Tannier X., Vilnat A., « Automatically Generated Noun Lexicons for Event Extraction », *Proceedings of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (CicLing 2012)*, New Delhi, India, March, 2012a.

Arnulphy B., Tannier X., Vilnat A., « Event Nominals : Annotation Guidelines and a Manually Annotated Corpus in French », *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May, 2012b.

Besançon R., Ferret O., Jean-Louis L., « Construire et évaluer une application de veille pour l'information sur les événements sismiques. », *in* G. Pasi, P. Bellot (eds), *Actes de la conférence CORIA*, Éditions Universitaires d'Avignon, p. 287-294, 2011.

Bethard S., Martin J. H., « Identification of Event Mentions and their Semantic Class », *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Sydney, Australia, p. 146-154, 2006.

Bittar A., Building a TimeBank for French : A Reference Corpus Annotated According to the ISO-TimeML Standard, PhD thesis, Université Paris 7 - École doctorale de Sciences du Langage, 2010.

Grover C., Tobin R., Alex B., Byrne K., « Edinburgh-LTG : TempEval-2 System Description », *Proceedings of the 5th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Uppsala, Sweden, p. 333-336, July, 2010.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. H., « The WEKA Data Mining Software : An Update », *SIGKDD Explorations*, 2009.

Kumar Kolya A., Ekbal A., Bandyopadhyay S., « JU_CSE_TEMP : A First Step towards Evaluating Events, Time Expressions and Temporal Relations », *Proceedings of the 5th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Uppsala, Sweden, p. 345-350, July, 2010.

Lafferty J., McCallum A., Pereira F., « Conditional random fields : Probabilistic models for segmenting and labe ling sequence data », *International Conference on Machine Learning (ICML)*, 2001.

Lavergne T., Cappé O., Yvon F., « Practical Very Large Scale CRFs », *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, Association for Computational Linguistics, p. 504-513, July, 2010.

Llorens H., Saquete E., Navarro B., « TIPSem (English and Spanish) : Evaluating CRFs and Semantic Roles in TempEval-2 », *Proceedings of the 5th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Uppsala, Sweden, p. 284-291, July, 2010.

Ney H., Essen U., Kneser R., « On Structuring Probabilistic Dependencies in Stochastic Language Modelling », *Computer Speech and Language*, vol. 8, p. 1-38, 1994.

Parent G., Gagnon M., Muller P., « Annotation d'expressions temporelles et d'événements en franąis », *in* F. Béchet (ed.), *Traitement Automatique des Langues Naturelles (TALN'08)*, Association pour le Traitement Automatique des Langues (ATALA), 2008.

Pustejovsky J., Castaño J., Ingria R., Saurí R., Gaizauskas R., Setzer A., Katz G., « TimeML : Robust Specification of Event and Temporal Expressions in Text », *IWCS-5, Fifth International Workshop on Computational Semantics.*, Tilburg University, 2003.

Pustejovsky J., Lee K., Bunt H., Romary L., « ISO-TimeML : An International Standard for Semantic Annotation », *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*, European Language Resources Association (ELRA), Valletta, Malta, May, 2010.

Pustejovsky J., Verhagen M., Saurí R., Littman J., Gaizauskas R., Katz G., Mani I., Knippen R., Setzer A., *TimeBank 1.2*, Linguistic Data Consortium. 2006.

Quinlan R., *C4.5 : Programs for Machine Learning*, Morgan Kaufman Publishers, 1993.

Saurí R., Knippen R., Verhagen M., Pustejovsky J., « Evita : A Robust Event Recognizer for QA Systems », *Proceedings of the HLT05*, Vancouver, Canada, OCT, 2005.

Tanguy L., Hathout N., « Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web », *in* J.-M. Pierrel (ed.), *Actes de Traitement Automatique des Langues Naturelles (TALN'02)*, vol. Tome I, ATILF, ATALA, Nancy, France, p. 245-254, June, 2002.

UzZaman N., Allen J., « TRIPS and TRIOS System for TempEval-2 : Extracting Temporal Information from Text », *Proceedings of the 5th International Workshop on Semantic Evaluation*, Association for Computational Linguistics, Uppsala, Sweden, p. 276-283, 2010.

UzZaman N., Llorens H., Derczynski L., Allen J., Verhagen M., Pustejovsky J., « SemEval-2013 Task 1 : TempEval-3 : Evaluating Time Expressions, Events, and Temporal Relations », *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2 : Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Association for Computational Linguistics, Atlanta, Georgia, USA, p. 1-9, 2013.

Verhagen M., Gaizauskas R., Schilder F., Hepple M., Katz G., Pustejovsky J., « SemEval-2007 Task 15 : TempEval Temporal Relation Identification », *Proceedings of the SemEval conference*, 2007.

Verhagen M., Saurí R., Caselli T., Pustejovsky J., « SemEval-2010 Task 13 : TempEval-2 », *Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010*, Uppsala, Sweden, p. 57-62, 2010.