# Supervised Machine Learning Techniques to Detect *TimeML* Events in French and English

Béatrice Arnulphy[1], Vincent Claveau[2](✉), Xavier Tannier[3], and Anne Vilnat[3]

[1] Inria - Rennes-Bretagne Atlantique, Rennes, France
beatrice.arnulphy@inria.fr
[2] IRISA-CNRS, Rennes, France
vincent.claveau@irisa.fr
[3] LIMSI-CNRS, University of Paris Sud, Orsay, France
{xavier.tannier,anne.vilnat}@limsi.fr

**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 recent years. However, 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 approach.

**Keywords:** Event identification · Information extraction · TimeML · TempEval · CRF · Language modeling · English · French

## 1 Introduction

Extracting events from texts is a keystone for many applications concerned with information access (question-answering systems, dialog systems, text mining...). During the last decade, this task received some attention through the *TempEval*[1] conference series (2007, 2010, 2013). In these challenges, participants were provided with corpora annotated with *TimeML* features (cf. Sec. 2.1) in several languages, as well as an evaluation framework. It allowed to obtain reference results and relevant comparison between event-detection systems.

Yet, despite the success of the multilingual *TempEval-2* challenge, no participant proposed systems for French, for any task. Up to now, the situation is such that:

- the few studies dealing with detecting events in French cannot be compared since they use different evaluation materials;
- the performance of the systems cannot be compared to state-of-the-art systems, mainly developed for English.

[1] http://www.timeml.org/tempeval2/.

C. Biemann et al. (Eds.): NLDB 2015, LNCS 9103, pp. 19–32, 2015.
DOI: 10.1007/978-3-319-19581-0_2

The work presented in this paper aims at addressing these two shortcomings by proposing several systems for detecting events in French. These systems are evaluated within different frameworks/languages so that they can be compared with state-of-the-art systems, in particular those developed for English. More precisely, the tasks that we are tackling are the identification of events and of nominal markers of events. The systems we propose are versatile enough to be easily adapted to different languages or data types. They are based on usual machine learning techniques – decision trees, conditional random fields (CRFs), k-nearest neighbors (kNNs) – but make use of lexical resources, either existing, or semi-automatically built. These systems are tested on different evaluation corpora, including those of *TempEval-2* challenge. They are applied to both English and French data sets; the English data allow us to assess their performance relative to other published approaches. Whereas the French data provide reference results for this language.

The paper is structured as follows: in Sect. 2, the context of this work is presented, including the *TempEval* extraction tasks and the TimeML standard. In Sect. 3, we propose a review of the state-of-the-art systems developed for these tasks. Our own extraction systems are then detailed (Sect. 4) and their results on English and French are respectively reported in Sects. 5 and 6.

## 2 Extracting Events: The TempEval Framework

The *TempEval* challenges offered a unique framework dedicated to event detection tasks. The tasks rely on the *TimeML* specification language. In the remaining of this section, we give insights into this standard and we detail the *TempEval* challenges.

### 2.1 TimeML

Event definition used in *TempEval* follows the *ISO-TimeML* language specification [21]. It was developed to annotate and standardize events and temporal expressions in natural language texts. According to this standard, an event is described in a generic way as *"a cover term for situations that happen or occur"* [20]. For instance, this annotation scheme considers[2]:

- event expressions (<EVENT>), with their class and attributes (time, aspect, polarity, modality). There are 7 classes of events: ASPECTUAL, I_ACTION, I_STATE, OCCURRENCE, PERCEPTION, REPORTING and STATE;
- temporal expressions and their normalized values (<TIMEX3>);
- temporal relations between events and temporal expressions (<TLINK>);
- aspectual ( <ALINK>) and modal (<SLINK>) relations between events;
- linguistic markers introducing these relations (<SIGNAL>).

This annotation scheme was first applied to English, and then to other languages (with small changes in the scheme and adaptations to the annotation

[2] For details and examples, see [23].

guide for each considered language). The *TimeML* annotated corpora are called TimeBank: *TimeBank 1.2* [19] for English, *FR-TimeBank* [7] for French, and so on. In practice, it is noteworthy that events in these corpora are mostly verbs and dates. Nominal events, though important for many applications, are less frequent, which may cause specific problems when trying to identify them (cf. Sects. 5 and 6).

In this article, we focus on identifying events as defined by the *TimeML* tag <EVENT> [29], which is the purpose of task B in *TempEval-2*. An example of such an event, from the TimeBank-1.2 annotated corpora[3], is given below: line 1 is the sentence with 2 events annotated, lines 2 and 3 describe the attributes of these events.

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

(2) <MAKEINSTANCE eventID=”e3” eid=”ei377” tense=”NONE” aspect=”NONE” po-. larity=”POS” pos=”NOUN”/>

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

### 2.2 TempEval Challenges

Up to now, there have been three editions of *TempEval* evaluation campaign (organized during *SemEval*[4]).

*TempEval-1*[5] [28] focused on detecting relations between provided entities. In this first edition, only English texts were proposed. *TempEval-2*[6] [29] focused on detecting events, temporal expressions and temporal relations. This campaign was multilingual (including English, French and Spanish) and the tasks were more precisely defined than for *TempEval-1*.

*TempEval-3*[7] [27] consisted again in the evaluation of event and temporal relation extraction, but only English and Spanish tracks were proposed. Moreover, a new focus of this third edition was to evaluate the impact of adding automatically annotated data to the training set.

As previously mentioned, in this paper, we mainly focus on extracting events (marked by verbs or nouns) as initially defined in *TempEval-2* challenge. Besides, as our goal is to produce and evaluate systems for French, we use the dataset developed for *TempEval-2* (as well as other French datasets that will be described below).

## 3 Related Work

Several studies have been dedicated to the annotation and the automatic extraction of events in texts. Yet, most of them were carried out in a specific framework,

[3] http://www.TimeBank-1.2/data/timeml/ABC19980108.1830.0711.html.

[4] http://semeval2.fbk.eu/semeval2.php.

[5] http://www.timeml.org/tempeval/.

[6] http://semeval2.fbk.eu/semeval2.php?location=tasks#T5.

[7] http://www.cs.york.ac.uk/semeval-2013/task1/.

with a personal definition of what could be an event. This is the case for example in monitoring tasks (for example on seismic events [11]), popular event detection from tweets [5] or in sports [14]. These task-based definitions of events are not discussed in this paper, as they often lead to dedicated systems and can hardly be evaluated in other contexts. In this section, we focus on the closest studies, either done within the *TempEval-2* framework or not, but relying on the generic and linguistically motivated definition of events as proposed in *TimeML*.

### 3.1 Extracting TimeML Events

Evita system [23] aims to extract *TimeML* events in *TimeBank1.2*, combining linguistic and statistical approaches, using *WordNet* as external resource. Step [6] aims at classifying every *TimeML* items with a machine learning approach based on linguistic features, without any external resources. They also develop two baseline systems (Memorize and a simulation of Evita). Although every *TimeML* elements were searched for, the authors focus specifically on nominal events. They reached the conclusion that the automatic detection of these events (*i.e.* nouns or noun phrases tagged <event>) is far from being trivial, because of the high variability of expressions, and consequently because of the lack of training data covering all the possible cases.

Parent et al. [18] worked on the extraction of *TimeML* structures in French. Their corpus of biographies and novels was manually annotated before *FR-TimeBank*'s publication. These studies primarily concern the adverbial phrases expressing temporal localization. Their model is mainly based on parsing and pattern matching of syntactic segments. Concerning nouns, they used their own reviewed version of the *VerbAction* lexicon [25] and few syntactic rules. To the best of our knowledge, this work is the only one concerning *TimeML* events on French.

### 3.2 Work Within Scope of TempEval-2

Several systems participated in *TempEval-2* campaign, most of them on the English dataset. The best ranked, TIPSem [16], learns CRF models from training data and the approach is focused on semantic information. The evaluation exercise is divided into four groups of problems to be solved. In the recognition problem group, the features are morphological (lemma, part-of-speech (PoS) context from *TreeTagger* [24]), syntactical (syntactic tree from *Charniak parser* [8]), polarity, tense and aspect (using PoS and handcrafted rules). The semantic level features are the semantic role, the governing verb of the current word, role configuration (for governing verbs), lexical semantics (the top four classes from WordNet for each word). This system being the best ranked of the challenge, it was later used as a reference for *TempEval-3*. Edinburgh [9] relies on text segmentation, rule-based and machine-learning named entity recognition, shallow syntactic analysis and lookup in lexicons compiled from the training data and from *WordNet*. Trips parser [1] provides event identification and "TimeML-suggested features", and is semantically motivated. It is based on a proper Logical Form Ontology. Trios [26] is based on Trips with a Markov Logic Network

(MLN) which is a Statistical Relation Learning Method (SRL). Finally, Ju_cse [12] consists in a very simple and manually designed rule-based method for event extraction, where all the verb PoS tags (from *Stanford* PoS tagger) are annotated as events.

All these systems and their respective performance provide valuable information. Firstly, most of them rely on a classical architecture using machine learning, and CRFs seem to perform well, as they do in many other information extraction tasks. Secondly, the results highlight the necessity of providing semantic information large enough to cover the great number of ways to express events, especially for the nominal events. The systems that we propose in this paper share many points with some of the systems we described here, as they also rely on supervised machine learning, including CRFs, and also make use of lexicons which were in part obtained automatically.

## 4 Event Detection Systems

The systems proposed in this paper aims at being easily adapted to any new language or text. To do so, as for many state-of-the-art systems, they adopt a supervised machine learning framework: *TimeML* annotated data are provided to train our systems, which are then evaluated on separate test set. The goal of the classifier is to assign each word with a label indicating whether it is an event. Since some events are expressed through multi-word expressions, the IOB annotation scheme is used (B indicates the beginning of an event, I is for inside an event, and O is for outside – if the word does not refer to an event). The training data are excerpts from corpora where each word is annotated with these labels and is described by different features (detailed hereafter). These data are then exploited by machine learning techniques presented in Sub-sects. 4.2 and 4.3. After the training phase, the inferred classifiers can be used to extract the events from unseen texts by assigning the most probable label to each word with respect to its context and features.

### 4.1 Features

The features used in our systems are simple and easy to extract automatically. They include what we call hereafter internal features: word-form, lemmas and part-of-speech, obtained with TreeTagger[8]). On the other hand, external features bring lexical information coming from existing lexicons, either general or specific to event description:

- for French, a feature indicates for each word whether it belongs to the *VerbAction* [25] and *The Alternative Noun Lexicon* [7] lexicons or not. The former lexicon is a list of verbs and their nominalization describing actions (*e.g. enfumage* (act of producing smoke), *réarmement* (rearmament)); the latter is complementary as it records non deverbal event nouns (nouns that are not derived from a verb, eg. *miracle* (miracle), *tempête* (storm)).

[8] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger.

- for English, a feature indicates for each word whether it belongs to one of the eight classes of synsets concerned with actions or events, that is *change, communication, competition, consumption, contact, creation, motion, stative.*

We also exploit lexical resources that are automatically built, called *Eventiveness Relative Weight Lexicons* (ERW hereafter), following the seminal work or Arnulphy et al. [3]. These lexicons are lists of words associated with the probability that they express an event. In our case, they are built from newspaper corpora (AFP news wire for French and Wall Street Journal for English). We do not go into further details about the building of ERWs, they may be found in the previously cited reference. It is worth noting that these lexicons bring information on polysemic words. It means that, for instance, most of the entries may express an action, which is then relevant to extract, or the result of an action, which is not wanted (*e.g. enfoncement, décision* in French). Thus, these lexicons are not sufficient by themselves, but they bring valuable information to exploit with more complex method taking the context into account.

### 4.2 CRF and Decision-Tree Based Systems

We have considered two machine learning techniques usually used for this kind of tasks: conditional random fields (CRFs, for instance used by [16]), and decision trees (DTs) that have shown good performance in previous work [2].

Concerning the DTs, we use the WEKA [10] implementation of C4.5 [22]. The interest of DTs is their ability to handle different types of features: nominal (useful to represent part-of-speech for example), boolean (does a word belong to a lexicon), numeric (ERW values). In order to take into account the sequential aspect of the text, each word is described by its own features (cf. sec. 4.1) and those of the preceding and following words.

CRFs [13] are now a well-established standard tool for annotation tasks. Contrary to DTs, they inherently take into account the sequential dependencies in our textual data. But in contrast, most implementations do not handle numeric features. Thus, the ERW scale of values is splitted into 10 equally large segments and transformed into a 10-value nominal feature. In the experiment reported below, we use WAPITI [15], a fast and robust CRF implementation.

### 4.3 CRF-kNN Combined System

The two systems described above are quite common for information extraction. We propose here a more original system, still based on CRFs, but aiming at addressing some of their shortcomings. One of them is the fact that CRFs consider the sequential context in a very constrained way. A sequence introducing an event X, as in example 1 below, will be considered as different to example 2 due to the offset caused by the insertion of "*l'événement de*" or "*unexpectedly*". The event Y may thus be undetected, even though example 1, which seems similar, is in the training set.

1. "*c'est à cette occasion que s'est produit X ...*" / at the very moment, X happened
2. "*c'est à cette occasion que s'est produit l'événement de Y ...*"/ at the very moment, unexpectedly, X happened

Another issue with CRFs is that available implementations can hardly handle numeric features (like ERW values), or consider sets of synonyms.

To address these different limits, we join a kNN classifier to CRFs to help to label the potential events. CRFs are used as explained in the previous section, but all the possible labels with their probabilities are kept instead of only the most probable label. The kNN then compute a similarity between every candidates (every potential events found by the CRFs, regardless of their probability) and all the training instances.

In our case, this similarity is computed by using n-gram language modeling. It allows us to estimate a probability (written $P_{LM}$) for a sequence of words. More precisely, for each potential event found by the CRF, its class $C^*$ (event or not) is decided following its probability given by the CRF ($P_{CRF}(C)$), and the probabilities provided by language models on the event itself and on its left and right contexts (resp. candidate, cont$_L$ and cont$_R$). Language models (*i.e.* sets of estimated probabilities) are thus estimated for each class and each position (left or right) from the training data. This is done by counting n-grams occurring at the left and at the right of each event of the training set, and inside the event. These models are denoted $\mathcal{M}_C$, $\mathcal{M}_C^R$ and $\mathcal{M}_C^L$. Finally, the label decision is formalized as:

$$C^* = \operatorname*{argmax}_C P_{CRF}(C) * P_{LM}(\text{cont}_L|\mathcal{M}_C^L) * P_{LM}(\text{candidate}|\mathcal{M}_C) * P_{LM}(\text{cont}_R|\mathcal{M}_C^R)$$

In our experiments, we use bigram models for $\mathcal{M}_C^D$ and $\mathcal{M}_C^G$, and unigram models for $\mathcal{M}_C$; the right and the left context are 5 words long. Based on that, the similarity of the left contexts of examples 1 and 2 would be high enough to detect the event in example 2.

Moreover, one other interest of language models is that it makes it possible to take into account lexical information during the smoothing process. In order to prevent unseen n-grams from generating a 0 probability for a sequence, it is usual to associate a small but non zero probability to them. Several strategies are proposed in the literature [17]. In our case, we use a back-off strategy from unseen bigrams to unigrams and a Laplacian smoothing, as it is easy to implement, for unseen unigrams. One originality of our work is to use also smoothing to exploit the information in our lexicons. Indeed, a word unseen in the training data may be replaced with a seen word belonging to the same lexicon (or synset for WordNet). When several words can be used, the one that maximizes the probability is chosen. In every case, a penalty ($\lambda < 1$) is applied; formally, for a word $w$ unseen in the training data for a model $\mathcal{M}$, we have:

$$P(w|\mathcal{M}) = \lambda * max\{P(w_i|\mathcal{M}) \,|\, w_i, w \text{ is the same lexicon/synset }\}$$

Concerning the ERW values, they give information on the presence of the considered word inside the lexicons, *i.e.* may be interpreted as belonging values (absent words are scored 0) which are used to compute the penalty for the smoothing: the replacement penalty ($\lambda$) between one unseen word $w$ with a seen one $w_i$ is proportional to the difference between the values of these two words.

Combining these two systems makes the most of the CRF ability to detect interesting phrases, thanks to a multi-criterion approach (part-of-speech, lemmas), and of the language modeling to consider larger contexts and to integrate lexical information as a smoothing process.

## 5 Experiments on English

### 5.1 Settings

To evaluate our systems, the metrics we adopt are the same as for *TempEval-2*: precision (Pr), recall (Rc) and F1-score (F1). They are computed for the whole extraction tasks as well as on a subset of events known to be more difficult, specifically nominal events (events expressed as a noun or a phrase whose head is a noun), and stative nominal events.

Beside the overall performance of the systems, we want to assess the importance of the different features. Here, we report the results for some of the several combinations we tested, according to the type of features: internal and/or external (cf. Sect. 4.1). The configurations tested are:

1. with internal information only: the models only rely on word forms, lemmas and part-of-speech.
2. with both internal and external information;
3. this configuration is a variant of the preceding one, specific to the use of WordNet: the 8 classes of synsets are used as 8 binary features indicating the presence or absence of the word in the synset classes.

### 5.2 Results

Among all the tested system/feature configurations, Table 1 present the results of the best ones. For comparison purposes, we also report the results of TIPsem, Edinburgh, ju_cse, trios et trips obtained at *TempEval-2*.

On these English data, CRF approaches outperform the ones based on decision trees, especially for the nominal event detection. This is partly due to the fact that nominal events are rare: only 7 % of nouns are events while, for instance, 57.5 % of the verbs are events. This imbalance has a strong impact on DTs while CRFs are less sensitive to that. But more generally, for any system, the performance drops when dealing with nominal events (either with or without states). Here again, this is due to the scarcity of such events, which are therefore less represented in the training data, which in turn causes a low recall. This study also shows that the performances differ depending on the different feature combinations. It shed light on the importance of using lexical information for these tasks, which confirms the state of the art.

**Table 1.** Performance of the best system/feature combination on the *TempEval-2* English data set.

| Type of event | System | Pr | Rc | F1 |
|---|---|---|---|---|
| All events | 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 | 0.80 | 0.79 |
| | (3) DT | 0.73 | 0.71 | 0.72 |
| Nominal only | (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 |
| Nominal without states | (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 |

Last, our CRF-kNN system yields the best results, outperforming CRFs alone, DT or state-of-the-art systems. These results are promising as they only rely on features that are easy to extract from the text (*e.g.* PoS) or publicly available (*e.g.* WordNet). Thus, they are expected to be applicable to any language such as French (cf. next section).

# 6 Experiments on French

## 6.1 Dataset and Comparison to English

In contrast to English, few corpora are available to develop, evaluate and compare event extraction systems in French. Among them, the *TempEval-2* French corpus is supposed to be similar to its English counterpart in terms of genre and annotation. As for the English corpus, which was part of the *TimeBank1.2*, this French corpus is a part of the *FR-TimeBank*. In previous work [4], we also proposed an annotated corpus for French. As for *FR-TimeBank*, it is composed of newspaper articles, which makes it comparable in genre to *En-TempEval-2* corpus, but it is only annotated in non-stative nominal events (*TimeML* tag <EVENT class="OCCURRENCE" pos="NOUN">).

Several points are worth mentioning for a fair comparison with English results. Table 2 shows that the proportion of all events is comparable between the French and English *TempEval-2* corpora: about 2.6 by sentence. However a detailed analysis shows that there are more verbal events than nominal ones in *TempEval-2* corpora, but relatively more nominal events in both French corpus

**Table 2.** Comparison of English (ENG) and French (FRE) corpora with *TimeML* annotations.

| | | # sentences | # tokens | # events |
|---|---|---|---|---|
| ENG | *TempEval-2* | 2,382 | 58,299 | 6,186 |
| FRE | *TempEval-2* | 441 | 9,910 | 1,150 |
| FRE | corpus of [4] | 2,414 | 54,110 | 1,863 |

**Table 3.** Performance of the best feature/system configurations on the French corpora (*Fr-TempEval-2*, [4] and [18]).

| Corpus | Type of event | System | Pr | Rc | F1 |
|---|---|---|---|---|---|
| *TempEval-2* français | all events | (2) CRF-kNN | **0.87** | **0.79** | **0.83** |
| | | (2) CRF | 0.80 | 0.76 | 0.78 |
| | | (4) DT | 0.78 | 0.77 | 0.78 |
| | nominal only | (2) CRF-kNN | **0.69** | 0.60 | **0.64** |
| | | (2) CRF | 0.55 | 0.52 | 0.53 |
| | | (4) DT | 0.58 | **0.63** | 0.60 |
| | nominal without states | (2) CRF-kNN | **0.65** | **0.52** | **0.58** |
| | | (2) CRF | 0.53 | 0.46 | 0.50 |
| | | (4) DT | 0.57 | 0.49 | 0.53 |
| Corpus of [4] | nominal without states | (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 of [18] | all events | Parent et al | 0.625 | 0.777 | 0.693 |
| | nominal only | Parent et al | 0.547 | 0.537 | 0.542 |

than for English. Furthermore, the corpus of [4] contains more nominal events than *Fr-TempEval-2*; and about 90 % of nominal events are not states in *Fr-TempEval-2*, versus 80 % in *En-TempEval-2*.

### 6.2 Results on French

The feature combinations used for English have been tested; Table 3 reports the best performing model/feature configurations. For purposes of comparison, we also implemented a system proposed in a previous work [2] to serve as a baseline, which we note (4). This system also relies on DTs but uses features that are more difficult to obtain and thus less adaptable, namely a deep syntactic analysis, post-edited with manually-built rules. Finally, we also report the results published by [18] on their own corpus.

Overall, the CRF models perform as well as the technique proposed in [2], while using no syntactic information and hand-coded resources. Concerning the non-stative nominal events, the results are significantly better on the corpus

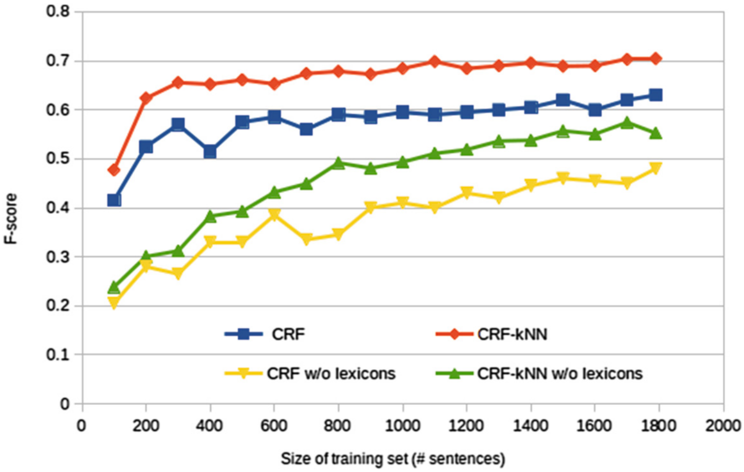

**Fig. 1.** Performance (F1-score) of CRF-kNN and CRF models with respect to the number of training sentences.

of [4] than on *Fr-TempEval-2* (F1=0.63 *vs.* F1=0.53). This performance gap highlights the above-mentioned intrinsic differences of the two corpora. Finally, even if the comparison is tricky since we deal with different corpora, it is worth noting that our systems outperform the results reported by [18].

French experiments lead to the same observations as for English data: extracting nominal events is more difficult than extracting verbal events. Yet, the difference between nominal and non-stative nominal events is smaller than for English. It may be explained by the proportion of such events which differs, as mentioned in Sect. 6.1. As for English, our system combining CRFs and language-model-based kNNs yields the best overall results. Again, the results obtained with the different sets of features underline the positive impact of lexical information for such extraction tasks.

### 6.3 Influence of Lexicons and Training Data Size

In order to evaluate the impact of the size of training data on the performance of our CRF-kNN system, we report in Fig. 1 how F1-score evolves according to the number of annotated sentences used for training. For purposes of comparison, we also report the performance of the CRF-alone system in order to shed the light on the contribution of the language models. Two configurations are tested: with and without external lexical information.

First, this figure shows that the interest of combining CRFs with the language-model kNNs is significant, for any size of the training data. Second, the language models improve the CRF performance, whether lexicons are used or not. Obviously, without external lexical information, the F-score progression depends directly on the number of training sentences. In contrast, using lexical resources makes the F1-score increase rapidly with small amount of training data, and then increase again linearly for bigger amount of data. It shows that small training set, and thus small annotation costs, can be considered, provided that lexical resources are available.

## 7 Conclusion

Extracting events from texts is a keystone for many applications, but definitions of what is an event are often *ad hoc* and difficult to generalize, which makes any comparison impossible. On the other hand, the linguistically motivated and standardized definition given by *TimeML* and implemented in the *TempEval* challenges was not completely explored for some languages such as French. In this paper, we tried to fill this gap by proposing several systems, evaluated on French, but also on English in order to assess their performance with respect to state-of-the-art systems.

The three proposed systems adopt a classical architecture based on supervised machine learning techniques. Yet, one of our contributions is to propose a combination of CRFs and language-model kNNs, which takes advantage of both techniques. In particular, the language model offers a nice way to incorporate lexical information in the event detection process, which has proven to be useful, especially when dealing with few data. This original combination of CRFs and kNNs yields good results on both English and French and outperforms state-of-the-art systems. The good results obtained for English validate our approach and suggest that the performance reported for French may now serve as a reasonable baseline for any further work. Among the perspectives, we will focus on the extraction of the other temporal markers and relations defined in *TimeML*. We also foresee the adaptation of our CRF-kNN method to these tasks as well as other information extraction tasks.

## References

1. Allen, J.F., Swift, M., de Beaumont, W.: Deep semantic analysis of text. In: Proceedings of the 2008 Conference on Semantics in Text Processing, STEP 2008, pp. 343–354. Association for Computational Linguistics, Stroudsburg (2008). http://dl.acm.org/citation.cfm?id=1626481.1626508
2. Arnulphy, B.: Désignations nominales des événements: Étude et extraction automatique dans les textes. Ph.D. thesis, Université Paris-Sud - École Doctorale d'Informatique de Paris Sud (EDIPS) / Laboratoire LIMSI (2012)
3. Arnulphy, B., Tannier, X., Vilnat, A.: Automatically generated Noun Lexicons for event extraction. In: Proceedings of the 13th International Conference on Intelligent Text Processing and Computational Linguistics (CicLing 2012), New Delhi, India, March 2012
4. Arnulphy, B., Tannier, X., Vilnat, A.: Event nominals: annotation guidelines and a manually annotated corpus in french. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey, May 2012

5. Becker, H., Naaman, M., Gravano, L.: Beyond trending topics: Real-world event identification on twitter. In: Fifth International AAAI Conference on Weblogs and Social Media (2011)
6. Bethard, S., Martin, J.H.: Identification of event mentions and their semantic class. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pp. 146–154. Association for Computational Linguistics, Sydney (2006). http://www.aclweb.org/anthology/W/W06/W06-1618
7. Bittar, A.: Building a TimeBank for French: a reference corpus annotated according to the ISO-TimeML standard. Ph.D. thesis, Université Paris 7 - École doctorale de Sciences du Langage (2010)
8. Charniak, E.: A maximum-entropy-inspired parser. In: 1st Meeting of the North American Chapter of the Association for Computational Linguistics, pp. 132–139 (2000), http://www.aclweb.org/anthology/A00-2018
9. Grover, C., Tobin, R., Alex, B., Byrne, K.: Edinburgh-ltg: Tempeval-2 system description. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 333–336. Association for Computational Linguistics, Uppsala, July 2010. http://www.aclweb.org/anthology/S10-1074
10. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. 11(1), 10–18 (2009)
11. Jean-Louis, L., Besançon, R., Ferret, O.: Text segmentation and graph-based method for template filling in information extraction. In: 5th International Joint Conference on Natural Language Processing (IJCNLP 2011), Chiang Mai, Thailand, pp. 723–731 (2011)
12. Kumar Kolya, A., Ekbal, A., Bandyopadhyay, S.: Ju_cse_temp: A first step towards evaluating events, time expressions and temporal relations. In: Proceedings of the 5th International Workshop on Semantic Evaluation. pp. 345–350. Association for Computational Linguistics, Uppsala, July 2010. http://www.aclweb.org/anthology/S10-1077
13. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labe ling sequence data. In: International Conference on Machine Learning (ICML) (2001)
14. Lanagan, J., Smeaton, A.F.: Using twitter to detect and tag important events in live sports. In: Artificial Intelligence (2011)
15. Lavergne, T., Cappé, O., Yvon, F.: Practical very large scale CRFs. In: Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 504–513. Association for Computational Linguistics, July 2010. http://www.aclweb.org/anthology/P10-1052
16. Llorens, H., Saquete, E., Navarro, B.: Tipsem (english and spanish): Evaluating crfs and semantic roles in tempeval-2. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 284–291. Association for Computational Linguistics, Uppsala, July 2010. http://www.aclweb.org/anthology/S10-1063
17. Ney, H., Essen, U., Kneser, R.: On structuring probabilistic dependencies in stochastic language modelling. Comput. Speech Lang. **8**, 1–38 (1994)
18. Parent, G., Gagnon, M., Muller, P.: Annotation d'expressions temporelles et d'événements en franąis. In: Béchet, F. (ed.) Traitement Automatique des Langues Naturelles (TALN 2008). Association pour le Traitement Automatique des Langues (ATALA) (2008)
19. 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). http://timeml.org/site/publications/timeMLdocs/timeml_1.2.1.html

20. Pustejovsky, J., Castaño, J., Ingria, R., Saurí R., Gaizauskas, R., Setzer, A., Katz, G.: Timeml: Robust specification of event and temporal expressions in text. In: IWCS-5, Fifth International Workshop on Computational Semantics, Tilburg University (2003)
21. Pustejovsky, J., Lee, K., Bunt, H., Romary, L.: ISO-TimeML: an international standard for semantic annotation. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010). European Language Resources Association (ELRA), Valletta (2010), http://aclweb.org/anthology-new/L/L10/
22. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufman Publishers (1993)
23. Saurí, R., Knippen, R., Verhagen, M., Pustejovsky, J.: Evita: a robust event recognizer for QA systems. In: Proceedings of the HLT 2005, Vancouver, Canada, October 2005
24. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of International Conference on New Methods in Language Processing, Manchester, UK (1994)
25. Tanguy, L., Hathout, N.: Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web. In: Pierrel, J.M. (ed.) Actes de Traitement Automatique des Langues Naturelles (TALN 2002), vol. Tome I, pp. 245–254. ATILF, ATALA, Nancy, France, June 2002
26. UzZaman, N., Allen, J.: Trips and trios system for tempeval-2: extracting temporal information from text. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 276–283. Association for Computational Linguistics, Uppsala (2010). http://www.aclweb.org/anthology/S10-1062
27. 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. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pp. 1–9. Association for Computational Linguistics, Atlanta, Georgia, USA (2013). http://www.aclweb.org/anthology/S13-2001
28. Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., Pustejovsky, J.: Semeval-2007 task 15: tempeval temporal relation identification. In: Proceedings of the SemEval Conference (2007)
29. Verhagen, M., Saurí, R., Caselli, T., Pustejovsky, J.: Semeval-2010 task 13: Tempeval-2. In: Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, Uppsala, Sweden, pp. 57–62 (2010). http://polyu.academia.edu/TommasoCaselli/Papers/1114340/TempEval2_Evaluating_Events_Time_Expressions_and_Temporal_Relations

Springer

http://www.springer.com/978-3-319-19580-3

Natural Language Processing and Information Systems
20th International Conference on Applications of Natural
Language to Information Systems, NLDB 2015, Passau,
Germany, June 17-19, 2015, Proceedings
Biemann, C.; Handschuh, S.; Freitas, A.; Meziane, F.; Métais,
E. (Eds.)
2015, XV, 460 p. 57 illus., Softcover
ISBN: 978-3-319-19580-3