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

Béatrice Arnulphy, Vincent Claveau, Xavier Tannier, Anne Vilnat

Supervised machine learning techniques to detect *TimeML* events in French and English

Béatrice Arnulphy*, Vincent Claveau†, Xavier Tannier§, and 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

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.

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

1 Introduction

Detecting events in 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). During these conferences, challenges were organized which consisted in providing participants with corpora annotated with *TimeML* features (cf. Sec. 2.1) in several languages, as well as an evaluation framework. It permitted to established reference results and to provide relevant comparison between event-detection systems.

Yet, despite the success of the multilingual *TempEval-2* challenge, no participant proposed systems for French, whatever the 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's ones (those developed for English for instance).

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

The work presented in this paper aims at addressing these two shortcomings by proposing system for detecting events in French. They 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 marks 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 (CRF), k-nearest neighbors (kNN) – but makes the most of lexical resources, either existing, or semi-automatically built. These systems are tested on different evaluation corpora, including those of *TempEval-2* challenge. In the one hand, they are applied to the English data set; it allows us to position their performance with respect to state-of-the-art's ones. In the other hand, they are applied to French data sets in order to provide reference results for this language.

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

2 Extracting events: the TempEval framework

The conferences *TempEval*, through the challenges that were organized, offered an unique framework dedicated to event detection tasks. The challenges chiefly rely on the language specification standard *ISO-TimeML*. In the remaining of this section, we first present some elements of this standard and then go into more detail about the *TempEval* challenges.

2.1 TimeML

The definition used in *TempEval* of what is a temporal event follows the *ISO-TimeML* language specification [20]. It was developed for annotating and standardizing events and temporal expressions in natural languages. According to this standard, an event is described in a generic way as *"a cover term for situations that happen or occur"* [19]. For instance, this annotation scheme makes it possible to marks in the texts (for details and examples, see [22]):

- event expressions (marked by a tag <EVENT>), with their class and attributes (time, spect, 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 and modal relations between events (respectively, <ALINK> and <SLINK>);

– linguisitic marks expressing these relations (<SIGNAL>).

This annotation scheme was first applied to English, and then to other languages (with small changes in the sceme and adaptations to the annotation guide for each considered language). The *TimeML* annotated corpora are called TimeBank: *TimeBank 1.2* [18] for English, *FR-TimeBank* [8] for French... In practice, it is noteworthy that events in these corpora are mostly verbs and dates. Nominal events, though important for many applications, are rarer, which may cause specific problems when trying to identify them (cf. Sections 5 and 6).

In this article, we are focusing on identifying events as defined by the *TimeML* tag <EVENT> [27]; this is equivalent to task B in *TempEval-2*. An example of such an event, from the TimeBank-1.2 annotated corpora[2], is given below: line 1 is the sentence with 2 events annotated, line 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 <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 TempEval challenges

Up to now, there had been three editions of *TempEval* evaluation campaigns (organized during *SemEval*[3]).
TempEval-1[4] [26] focused on detecting relations between provided entities. In this first edition, on ly English texts were proposed. *TempEval-2*[5] [27] 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[6] [25] was in the continuity of the preceding editions. Here again, it consisted in the evaluation of event and temporal relation extraction, but only English and Spanish tracks were proposed. Moreover, one new focus of this third edition was to evaluate the impact of adding to the training data set automatically annotated data in addition to the manually annotated ones.

As it was previously mentioned, in this paper, we chiefly focus on extracting events (marked by verbs or nouns) as initially defined in *TempEval-2* challenge. Beside, as our goal is to produce and evaluate systems for French, we use the data-set developed for *TempEval-2* (as well as other French data sets, see below).

[2] `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/`

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, with a personal definition of what could be an event. This is the case for example in monitoring tasks [6, for example on seismic events], popular event detection from tweets [5] or in sports [13]. These task-based definitions of events are not discussed in this paper, as they often lead to dedicated systems and are impossible to evaluate 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 Extrating TimeML events

Here, we mention work on event adopting the *TimeML* definition, in English and then in French. EVITA system [22] aims to extract *TimeML* events in *TimeBank1.2*, combining linguistic and statistical approaches, using *WordNet* as external resource. In STEP, [7] 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 (ie nouns or noun phrases tagged ¡event¿) is far from being trivial, because of the high variability of expressions, and consequently of the lack of training data covering all the possible cases.

[17] 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 of 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 [23] and few syntactic rules. To the best of our knowledge, this work is the unique one concerning *TimeML* events on French.

3.2 Work in the framework of TempEval-2

Several systems were proposed in the framework of *TempEval-2*, most of them working on English. The best one, TIPSEM [15] learns CRF models from training data and the approach is focused on semantic information. The evaluation exercise is divided in four groups of problem to be solved. In the recognition problem group, the features are morphological (lemma, Part-of-Speech (PoS) context from *TreeTagger*), syntactical (syntactic tree from *Charniak parser*), 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 of the challenge, it was later used as reference for *TempEval-3*. EDINBURGH [9] relies on text segmentation, rule-based and machine-learning named entity recognition and 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 [24] is based on TRIPS with a Markov Logic Network (MLN) which is a Statistical Relation Learning Method (SRL). Finally, JU_CSE [11] 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 performance provide important lessons. First, most of them rely on a classical architecture using machine learning, and unsurprisingly, CRF seem to perform well, as for other information extraction task. Secondly, the results highlight the necessity of providing semantic information large enough to cover the wide variety of expressing events, especially for the nominal ones. The systems that we propose in this paper share many common points with the state-of-the-art's ones (cf. next section) as they also rely on supervised machine learning, including CRF, and also rely on lexicons 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 s separate test data set. The task which is learned is a text annotation one: the goal of the classifier is to assign each word with a label that indicates whether it is an event or not. 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, that if the word do not refer to an event). The training data are corpora excerpts annotated with these labels for each word and are described by different features (detailed hereafter). These data are then exploited by machine learning techniques presented in sub-section 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 their context and features.

4.1 Features

The features used in our systems are simple and easy get automatically. They include what we call hereafter internal features: word-form, lemmas and part-of-speech, obtained with TreeTagger (`http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger`). Other features, said external, bring lexical information that seem important for our extraction tasks. They partly come from existing lexicons, either generalist or specialized on event description:

- for French, a feature indicates for each word whether it belongs to the *VerbAction* [23] and *The Alternative Noun Lexicon* [8] lexicons or not. The former lexicon is a list of verb and their nominalization describing actions (eg. *enfumage* (act of putting smoke), *réarmement* (rearmament); the latter is complementary as it records non deverbal event nouns (nouns that are non derived from a verb, eg. *miracle* (miracle), *tempête* (storm)).
- 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 their probability to express an event. In our case, they are built from newspaper corpora (AFP news feed for French and Wall Street Journal for English). We do not go into further details about the building of ER, they may be found in the previous 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 (eg. *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

As it was previously mentioned, the event detection task is seen here as an annotation one for which we train a classifier from annotated data. We have considered two machine learning techniques usually used for this kind of task: conditional random fields (CRF) [15, for instance used by], and decision trees (DT) that have shown good performance in previous work [2].

Concerning the DT, we use the WEKA [10] implementation of C4.5 [21]. The interest of DT is their ability to handle different type 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 features (cf. sec. 4.1) and those of the preceding and following words.

CRF [12] are now a well-establish standard tool for annotation tasks. Contrary to DT, 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 split into 10 equally large segments and transformed into a 10-value nominal feature. In the experiment reported below, we use WAPITI [14], a fast and robust CRF implementation.

4.3 CRF-kNN combined system

While the two previous systems are relatively usual for information extraction, we propose here another system, still based on CRF, but aiming at addressing some of its shortcomings. For instance, CRF consider the sequential context, but with in a very constrained way. A sequence introducing an event X, as in Example 1 below, will be considered as different than 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

Other shortcomings of CRF are that it is difficult to handle numeric features (like our ERW values) with the available implementations, or to indicate possible synonyms.

To address these different limits, we join a kNN classifier to CRF to help label the potential events. CRF is 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 candidate (every potential events found by the CRF, whatever 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 every potential event found by the CRF, its class C^* (event or not) is decided based on its probability given by the CRF ($P_{CRF}(C)$), and the probabilities provided by language models on the event itself and its left and right contexts (resp. candidat, cont_L and cont_R) Language models (i.e. sets of probabilities estimated) 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 right of each event of the training set, and inside the event. These models are noted $\mathcal{M}_C$, $\mathcal{M}_C^R$ and $\mathcal{M}_C^L$. Finally, the label decision is formalized as:

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

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 size of the right and left context is 5 words. Based on that, the similarity of the left contexts of Examples 1 and 2 would be high enough to detect the event of Example 2.

Moreover, one other interest of language models is that it makes it possible to integrate lexical information through the smoothing process. In order to prevent unseen n-grams to generate 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 [16]. In our case, we use a back-off strategy from unseen bigrams to unigrams and a Laplacian smoothing, 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 (λ ¡ 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 are directly interpreted as belonging values (absent words are scored 0) which are used to compute the penalty for the smoothing: the replacement penalty between one unseen word with a seen one 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 Setting

To evaluate our systems, we adopt the same scores than for *TempEval-2*, that is Precision (Pr), recall (Rc) and F1-score (F1). They are computed for the whole extraction tasks, but also 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 nominal events but states.

Beside the performance of the systems, we also 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 used: internal and/or external (cf. section 4.1). The configurations tested are:

1. with internal information and no external one: the models only rely on word-forms, lemmas and Part-of-Speech.
2. with 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 of absence of the word in the synsets.

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*.

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	0.8	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

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

On these English data, CRF approaches outperform the ones based on Decision Tree, specially for the nominal event detection. This results is due in part 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 DT while CRF are less sensitive to that. But more generally, whatever the system, one can observe a performance drop 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. The differences of performance of the feature combinations also shed light on the importance of using lexical information for these tasks. It was already supposed by the state-of-the-art review but is now confirmed by comparing systems that are completely identical beside the feature sets.

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

6 Experiments on French

6.1 Data set and comparison to English

In contrast with 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 news papers, which makes it comparable in genre to *En-TempEval-2* corpus, but is only annotated in non state nominal events (thus it corresponds the *TimeML* tag ¡EVENT class="OCCURRENCE" pos="NOUN"¿)

Several points are worth mentioning to fairly compare the results obtained on these corpora with the English ones. Table 2 presents some figures about the corpora. We can observe that the proportion of all events is comparable between the French and English *TempEval-2* corpora: about 2.6 by sentence. Moreover 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 corpora 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*.

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

6.2 Results on French

The same feature combinations 1 and 2 used for English have been tested; Table 3 report the best performing model/feature configurations. For comparison purposes, 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 DT but use feature that are more difficult to obtain and thus less adaptable, namely a deep syntactic analysis, post-edited with manually-built rules. Last, we also report the results published by [17] 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-state nominal events, the results are significantly better on the corpus of [4] than on *Fr-TempEval-2* (F1=0.63 vs. F1=0.53). This performance gap highlights the intrinsic differences of the two corpora that were previously mentioned. Last, even if the comparison is delicate since we are dealing with different corpora, it is worth noting that our systems outperform the results reported by [17].

The same tendencies then for English are observed fro French: extracting nominal events is more difficult than dealing with any type of events. Yet, the difference between nominal and non state nominal events is smaller than for English. It may be explained by the difference of proportion of such events mentioned in Section 6.1. As for English, our system combining CRF and language-

Corpus	Type of event	System	Pr	Rc	F1
TempEval-2 français	all events	(2) CRF-kNN	**0.87**	**0.79**	**0.83**
		(2) CRF	0.8	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**	0.6
	nominal without states	(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
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 [17]	all events	Parent et al.	0.625	0.777	0.693
	nominal only	Parent et al.	0.547	0.537	0.542

Table 3. Performance of the best feature/system configurations on the French corpora (*Fr-TempEval-2*, [4] and [17]).

model based kNN yields the best results overall. It is also noteworthy that the results obtained with the different sets of features underline the positive impact of lexical information for such extraction tasks.

6.3 About the influence of lexicons and training set size

In order to evaluate the impact of the size of training data on the performance of our CRF-kNN system, we report in Figure 1 how F1-score evolves according to the number of annotated sentences used for training. For comparison purposes, 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 CRF with the language-model kNN is significant, whatever the size of the training data. Second, the language models improve the CRF performance, when lexicons are used or not. Obviously, without external lexical information, the F-score progression depends directly of the number of training sentences. In contrast, using lexical resources makes the F1-score increasing rapidly with small amount of training data, and then is linear again for bigger amount of data. It shows that small training set, and thus small annotation costs, can be foreseen provided that lexical resources are available.

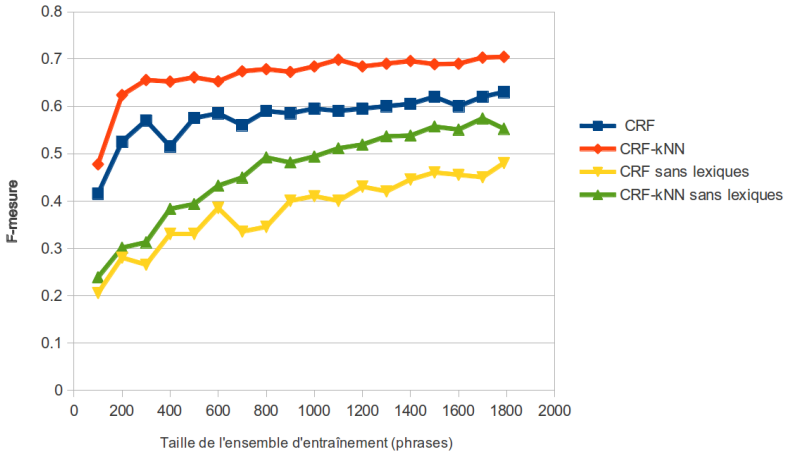

Fig. 1. Performance (F1-score) of CRF-kNN and CRF models with respect to the number of training sentences.

7 Conclusion

In the one hand, extracting events from texts is a keystone for many applications, but ad hoc definitions of what is an event are often employed, which make 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 like French. In this paper, we tried to fill that 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's systems.

The three proposed systems adopt a classical architecture based on machine learning techniques. Yet, one of our contributions is to propose a combination of CRF and language-model kNN handling, which makes it possible to take advantage of both techniques. In particular, the language model offer a nice way to incorporate lexical information in the event detection process, which we shown to be useful, especially when dealing with few data. This original combination of CRF and kNN 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. pp. 343–354. STEP '08, Association for Computational Linguistics, Stroudsburg, PA, USA (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 (Mar 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'12). 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. Besançon, R., Ferret, O., Jean-Louis, L.: Construire et évaluer une application de veille pour l'information sur les événements sismiques. In: Pasi, G., Bellot, P. (eds.) Actes de la conférence CORIA. pp. 287–294. Éditions Universitaires d'Avignon (2011)
7. 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, Australia (2006), http://www.aclweb.org/anthology/W/W06/W06-1618
8. 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)
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, Sweden (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 Explorations 11(1) (2009)
11. 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, Sweden (July 2010), http://www.aclweb.org/anthology/S10-1077
12. 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)
13. Lanagan, J., Smeaton, A.F.: Using twitter to detect and tag important events in live sports. Artificial Intelligence (2011)
14. 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

15. 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, Sweden (July 2010), http://www.aclweb.org/anthology/S10-1063
16. Ney, H., Essen, U., Kneser, R.: On structuring probabilistic dependencies in stochastic language modelling. Computer Speech and Language 8, 1–38 (1994)
17. 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'08). Association pour le Traitement Automatique des Langues (ATALA) (2008)
18. 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
19. 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)
20. 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'10). European Language Resources Association (ELRA), Valletta, Malta (May 2010), http://aclweb.org/anthology-new/L/L10/
21. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufman Publishers (1993)
22. Saurí, R., Knippen, R., Verhagen, M., Pustejovsky, J.: Evita: A Robust Event Recognizer for QA Systems. In: Proceedings of the HLT05. Vancouver, Canada (OCT 2005)
23. 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'02). vol. Tome I, pp. 245–254. ATILF, ATALA, Nancy, France (Jun 2002)
24. 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, Sweden (2010), http://www.aclweb.org/anthology/S10-1062
25. 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
26. 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)
27. 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. pp. 57–62. Uppsala, Sweden (2010),

http://polyu.academia.edu/TommasoCaselli/Papers/1114340/TempEval2_
Evaluating_Events_Time_Expressions_and_Temporal_Relations