Detecting salient events in large corpora by a combination of NLP and data mining techniques

Delphine Battistelli, Thierry Charnois, Jean-Luc Minel, Charles Teissèdre

Detecting Salient Events in Large Corpora by a Combination of NLP and Data Mining Techniques

Delphine Battistelli[1], Thierry Charnois[2,3], Jean-Luc Minel[2], and Charles Teissèdre[1]

[1] STIH, Université Paris Sorbonne, France

[2] GREYC, Université de Caen, France

[3] MoDyCo, UMR 7114, Université Paris Ouest Nanterre La Défense, France

delphine.battistelli@paris-sorbonne.fr, thierry.charnois@unicaen.fr,
ean-luc.minel@u-paris10.fr, charles.teissedre@gmail.com

Abstract. In this paper, we present a framework and a system that extracts "salient" events relevant to a query from a large collection of documents, and which also enables events to be placed along a timeline. Each event is represented by a sentence extracted from the collection. We have conducted some experiments showing the interest of the method for this issue. Our method is based on a combination of linguistic modeling (concerning temporal adverbial meanings), symbolic natural language processing techniques (using cascades of morpho-lexical transducers) and data mining techniques (namely, sequential pattern mining under constraints). The system was applied to a corpus of newswires in French provided by the *Agence France Presse* (AFP). Evaluation was performed in partnership with French newswire agency journalists.

Keywords. Dates, temporal adverbials, event extraction, sequential pattern.

Detección de destacados eventos en un corpus grande combinando técnicas para PLN y minería de datos

Resumen. En este trabajo se presenta el marco y el sistema para extracción de los eventos "destacados" relevantes a una pregunta de una gran colección de documentos, el cual también permite ubicar los eventos a lo largo de la línea de tiempo. Cada evento se representa por una frase extraída de la colección. Se han realizado unos experimentos que muestran el interés del método para este problema. El método propuesto se basa en la combinación del modelado lingüístico (con respecto a significados adverbiales temporales), las técnicas simbólicas de procesamiento de lenguaje natural (usando cascadas de transductores morfo-léxicos) y técnicas de minería de datos (la minería de patrones secuenciales bajo restricciones). El sistema ha sido aplicado a un corpus de noticias en idioma francés proporcionado por la *Agencia France Presse* (AFP). La evaluación se realizó en colaboración con periodistas de agencias francesas de noticias.

Palabras clave. Fechas, adverbiales temporales, extracción de eventos, patrón secuencial.

1 Introduction

A very big size of many document collections (newswires, RSS feeds, blogs, etc.) has led some users to express needs in tools that help them to have an organized and compact view of the information based on certain criteria, for example to have an overview of the most important events reported. While the definition of "importance" is subjective, events that are of interest to many people are often reported in many different documents (from different sources); updates and commentaries on these events also persist for a period of time. Our system makes use of this property to determine whether events are "important". More precisely, we assume that the more a given date is mentioned in a collection of documents, the more the event/s which is/are linked to this date is/are likely to be important. Our methodology is based on this assumption. The system which implements it first extracts salient dates, and then extracts the sentences in which those dates are mentioned (namely, extracts events which are linked to the dates). As a result, our system is able to summarize a large collection of documents (that could have been returned by a query-based search) by placing sentences that

report “important” events related to the query along a timeline. The experiments we have conducted show the interest of our methodology for this problem.

In this paper, we describe our framework in detail. In our experiments, we used a corpus of newswires written in French provided by *Agence France Presse* (AFP) for the development and evaluation of the system. This kind of corpus, by its very nature, contains textual redundancies concerning current events which are regularly updated and commented on. Our approach combines NLP symbolic techniques (to process temporal adverbial meanings in texts) and data mining techniques (namely, sequential pattern mining under constraints) to extract the set of sentences that report important events from texts. Because we use an unsupervised method, this makes the method generic and suited for all kinds of large (and redundant) corpora.

In what follows, we first review some of the related work in Section 2. Section 3 gives an overview of the system and presents the NLP and data mining tools used; we also describe how, when combined, they allow us to extract some salient events. Experiments, results and the evaluation are presented in Section 4.

2 Related Work

Our work aims at developing a temporal analysis protocol for extracting salient events from very large corpora. From the temporal point of view, it is closely related to two main kinds of studies: multidocument summarization (MDS) systems using temporal criteria, and systems to build textual and graphical timelines. From the event point of view, it is closely related to what is generally referred to as Event Extraction, but, to the best of our knowledge, no approach has yet been proposed to combine data mining and symbolic NLP techniques to automatically extract events from texts. However, it is clear that all these domains tend to overlap.

Previous work on MDS focused on summarizing similarities and differences in small clusters of documents (the summarization tasks as defined by the Document Understanding Conferences (DUC) require summarizing clusters of around ten documents) [1, 2]. For MDS, temporal processing enables a system to detect redundant excerpts from various texts on the same topic and to present results in a relevant chronological order [3]. The area of Topic Detection and Tracking (TDT) produced some work on the summarization of incoming streams of information. Most of these systems are based on statistical bag-of-words models that use similarity measures to determine proximity between documents; [4, 5] extracted associations between date and location and showed that the detected associations corresponded well with known events. [6] is a pioneer work on timeline overviews. The authors extracted noun phrases and named entities salient in specific time periods, and attempted to organize them into stories by grouping the phrases into clusters. However, the user still has to decipher the story from these clusters. [7] presented a system that uses measures of pertinence and novelty to construct timelines that consist of one sentence per date. While their summary is temporal in that they worked on incoming streams of articles, and aimed at including in a summary sentences containing useful and novel information, their work did not deal with summarizing multiple events on a timeline. [8] proposed a system that extracts events relevant to a query from a large collection of documents. Important events are those reported in a large number of news articles and each event is constructed according to one single query and represented by a set of sentences. More recently, [9] used a summarization-based approach to automatically generate timelines, taking into account the evolutionary characteristics of news. In order to automatically build event timelines from a search query, [10] present an approach which also detects more frequently reported dates in texts.

To the best of our knowledge, however, no approach has yet been proposed to summarize a large collection of documents (that could have been returned by a query-based search) by extracting and placing along a timeline sentences that report important events related to the query with only one sentence per date. This task is made possible by the use of a linguistically motivated approach to the meanings of temporal adverbials and the use of data mining techniques.

Some works (for instance [11]) takes advantage of the symbolic nature of the patterns discovered by data mining to learn linguistic patterns for named entity relation extraction. However, our work is a first attempt to use sequential data mining techniques to extract events from texts.

3 Methodology

3.1 Basic Assumptions and General Workflow

Two main principles underlie our methodology and allow us to use data mining techniques for extracting salient events from texts. The first one, already mentioned in the introduction, is a quantitative one: it says that the more a date is mentioned in a collection of documents, the more important it is likely to be. The second one is a qualitative one and relies on linguistic evidence: it says that, in the context of a sentence containing a date, the textual content is able to describe event(s) related to this date if, and only if, this date (e.g. *3rd March 2012*) appears in a temporal locating adverbial of type In Date (e.g. *during/on/... the 3rd March 2012*) –in contrast to a temporal locating adverbial such as *after/before/since/... the 3rd March 2012*. This second principle is of crucial importance: it bypasses the two –related– complex problems of syntactic attachments of adverbials and temporal relations between dates and events; combined with the first principle, it makes possible to use data mining techniques in order to extract the most informative sentence unit in which this date appears.

--2012

-JANUARY
-5: At least 68 killed in the anti-Shiite attacks in Baghdad and Nassiriya (South) (*Au moins 68 morts dans des attentats antichiites à Bagdad et Nassiriya (sud)*).
-FEBRUARY
-19: A suicide attack in front of the police academy of Baghdad claimed by the “Islamic state” of Irak (ISI). 15 killed, 21 wounded. (*Attentat suicide devant l'académie de police de Bagdad, revendiqué par l'"Etat islamique en Irak" (ISI): 15 morts, 21 blessés*)

Fig. 1. AFP hand-produced chronology of events (extract)

Regularly during the year, AFP journalists are required to propose “chronologies” (textual event timelines) in order to contextualize current events. These hand-produced chronologies may concern any topic discussed in the media, and consist in a list of dates (typically between 10 and 20) associated with a text describing the related event/s. Fig. 1 gives some translated examples of such a chronology.

In order to automatically build this kind of handmade chronology, we design the following workflow (Figure 2). Considering a thematic perspective and a given timespan, we first filter in the corpus collection a sub-corpus corresponding to the given theme (keyword based filtering). We

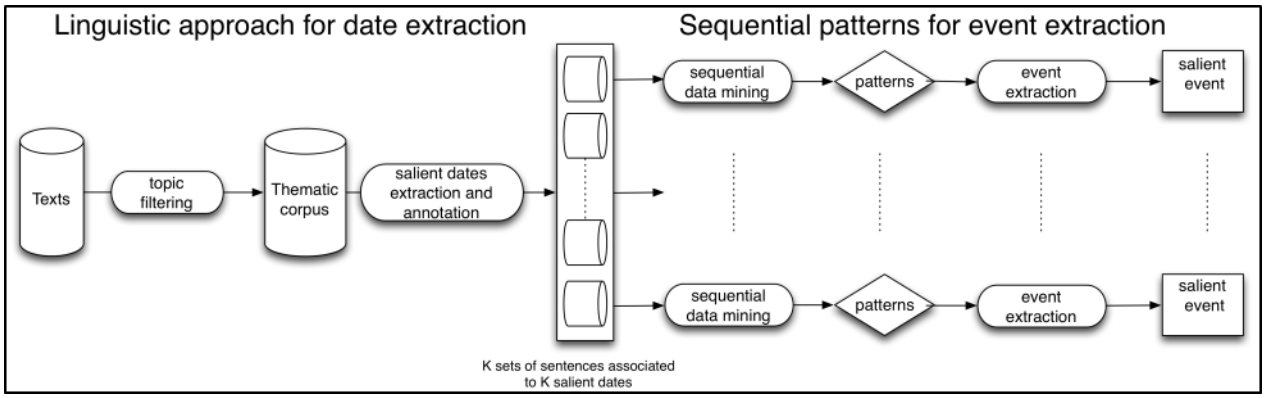

Fig. 2. Framework for event extraction

then extract salient dates thanks to rule-based techniques that provide a semantic description of temporal locating adverbials found in texts. The adverbials are then normalized: a calendar value is associated to calendar or deictic adverbials.

The set of dates extracted is filtered on three criteria in order to provide the K most salient dates: each extracted date must be included in the timespan considered, it must be part of the K most frequent dates set and it must suit a semantic filter that considers only adverbials that anchor an utterance inside the period of time they refer to. A set of sentences is associated to each salient date. The system then performs salient event extraction: firstly, sequential mining is used to discover frequent patterns under constraints, and secondly, the patterns enable sentence retrieval techniques to provide the sentence associated to a salient event. In the rest of this section, we describe the linguistic approach to detect dates and select salient dates. The sequential pattern extraction and the sentence retrieval process are then described in the next sub-section.

3.2 A Linguistic Approach to Date Extraction

Our approach to temporal expression annotation differs from the TimeML standard generally used in such tasks [12], in that it does not annotate a "date" (as defined in [12] with label TIMEX, e.g. *3rd March 2012, yesterday*), but rather a temporal locating adverbial (adjunct). This textual unit shares a semantic function and allows to analyze in a unified manner different kinds of temporal expressions [13, 14] generally analyzed separately (such as in [12]). Indeed, this approach establishes a distinction between a temporal basis (this temporal basis can be either a date or an event) and the temporal locating function that is expressed by a set of operators interacting with a temporal basis as in following examples (the sets of operators are underlined): *on the 3rd March 2012, since the end of the 30s, since the end of the election campaign*. An adverbial such as "*two days before the election*" is thus considered as a whole semantic unit and our approach consists in describing the semantic relationship between the "duration" ("two days"), the "signal" ("before") and the "event" ("election"), to use TimeML terminology. According to TimeML standard, such expression would either not lead to an overall analysis or would lead to an analysis where the duration would be linked to the event thanks to a temporal link ("before"), which is semantically inconsistent: *two days before* should indeed be analyzed as a complex operator. This unified analysis also allows to take into account various time periods, such as fuzzy periods (e.g. *by the end of the month, by the end of the Mesozoic era*). This is why, in the annotation process, we also distinguish the semantic description task and the "normalization" task.

Table 1. SDB1, a sequence database

SequenceID	Sequence
1	⟨*a b c d*⟩
2	⟨*b d e b*⟩
3	⟨*a c a*⟩
4	⟨*a d c*⟩

3.3 Sequential Patterns for Event Extraction

The second step of our approach aims at extracting salient events, using a sequential data mining approach. Recall that our assumption is that salient events are frequent in our corpus. Moreover, data mining makes it possible to find regularities in a large database in the form of patterns. Based on these ideas, our proposition consists in applying data mining on texts, specifically sequential data mining so as to take natural word order into account. We consider the corpus as the sequential database. Note that unlike machine learning based approaches, the training corpus is raw and has no annotated relations (such as syntactic relations, event-actor relations, etc.); the process is thus unsupervised. Once the sequence database has been built, sequential patterns under constraints are extracted.

Sequential Patterns. Sequential pattern mining is a data mining technique introduced in [15] to find regularities in a sequence database. There are several algorithms to extract sequential patterns [16, 18].

A *sequence s* is an ordered list of literals called *items*, denoted by $s = \langle i_1...i_m \rangle$ where $i_1...i_m$. are items. A sequence $S_1=\langle i_1...i_n \rangle$ is included in a sequence $S_2=\langle i'_1...i'_m \rangle$ if there exist integers $1 \leqslant j_1 <...< j_n \leqslant m$ such that $i_1 = i'_{j1},..., i_n = i'_{jn}$. S_1 is called a *subsequence* of S_2. S_2 is called a *supersequence* of S_1. It is denoted by $S_1 \leqslant S_2$. For example the sequence $\langle a\ b\ c\ d \rangle$ is a *supersequence* of $\langle b\ d \rangle$: $\langle b\ d \rangle \leqslant \langle a\ b\ c\ d \rangle$. A sequence database *SDB* is a set of tuples(S_{id}, S), where S_{id} is a sequence identifier and S a sequence. For instance, Table 1 depicts a sequence database of four sequences. A tuple *(Sid, S)* contains a sequence S_1, if $S_1 \leqslant S$. The support of a sequence S_1 in a sequence database *SDB*, denoted $sup(S_1)$, is the number of tuples in the database containing S_1. For example, in *SDB1* $sup(\langle b\ d \rangle) = 2$, since sequences 1 and 2 contain $\langle b\ d \rangle$. A frequent sequential pattern is a sequence such that its support is greater or equal to the support threshold: *minsup*. The sequential pattern mining extracts all the regularities which appear in the sequence database.

The set of frequent sequential patterns can be very large. The constraint-based pattern paradigm [19] provides useful techniques to express a user's interest in order to focus on the most promising patterns. A very widespread constraint is the frequency. However, it is possible to define many other useful constraints. For event extraction, we use three constraints: the gap constraint, the min length and the maximality.

A sequential pattern with a gap constraint [M, N], denoted by P[M,N], is a pattern such that at least M items and at most N items are allowed between every two neighbor items in the matched sequences. For instance, in Table 1, P[0,1] = ⟨(a)(c)⟩ and P[0,0] = ⟨(a)(c)⟩ are two patterns with gap constraints. P[0,1] matches three sequences (1, 3 and 4) whereas P[0,0] matches only one sequence (3), since in Sequences 1 and 4, there is one item between the item *a* and the item *c*. The min length constraint is useful to remove sequential patterns that are too generic and too small with respect to the number of items (number of words) to be relevant linguistic patterns. The max constraint is used in order to eliminate redundancy: a frequent sequential pattern *S* is maximal if there is no other frequent sequential pattern *S'* such that $S \leqslant S'$.

4 Experimentation, Results and Evaluation

4.1 Corpus

For our experiments, we used a corpus of newswires provided by the French news agency AFP. It is composed of 309 000 documents that span the January-October 2012 period. Each document is an XML file containing a title, a date of creation (DCT), a set of keywords, and textual content split into a few paragraphs. Editorial instructions are clear and concern the ability to identify very quickly who, where, when and what. This kind of corpus describes current events (and therefore contains a very large proportion - of the order of 40% - of deictic temporal expressions such as "*ce matin*" (*this morning*), "*aujourd'hui*" (*today*)) and also textual redundancies because descriptions of current events are regularly updated. There are only about 7% of absolute temporal references, and when present, they are frequently used to refer to a historical event which contextualizes the description of the current event (e.g. "*depuis la chute de la monarchie en Egypte en 1952*" (*since the collapse of the monarchy in Egypt in 1952*); "*depuis les accords de paix israélo-égyptiens de 1979*" (*since the Israelo-Egyptian peace agreement of 1979*)).

Sub-corpora. To extract salient events related to a given topic, we first filter a sub-corpus using keywords describing the topic. For our experiments, we filtered four sub-corpora whose topics were related to those of the human chronologies already produced by AFP journalists:

1. A set of newswires related to Egypt in order to compare our results with the AFP chronology entitled "*Egypt since the election of Mohamed Morsi*" (covering the period of time running from May 23rd to August 27th 2012). The set contained 1211 newswires.
2. A set of newswires related to the French car manufacturer PSA entitled "*PSA: restructuring plan chronicle*" (February 15th to July 23rd 2012). The set contained 849 newswires.
3. A set of newswires related to Irak: "*Irak: main attacks since the withdrawal of American soldiers at the end of 2011*" (June 9th 2011 to

July 23rd 2012). We only considered a subset of the chronology covering the year 2012. This subset contained 347 newswires.

4. A set of newswires related to Sudan; "*Sudan: a month of demonstrations against the regime*" (June 16th to July 13th 2012). The set contained 50 newswires.

4.2 Salient Dates Identification

The first criterion is quantitative and is based on the hypothesis that the more a given date is mentioned in a text corpus, the more likely it is to be associated with an important event, an event that is worth being repeatedly mentioned. Indeed, in each sub-corpus, we evaluate the salience of a date by the frequency with which it is mentioned. In the set of extracted dates, we consider those included in the period of time that the user is interested in. First we consider K salient dates, where K is a parameter that corresponds to the number of events related to a topic that the user decides to take into account. In this step, the system produces a list of sentence clusters C_{i1} linked to salient dates D_i (i=1...K) ordered by frequency.

The date extraction step is based on a temporal locating adverbial annotator presented in [14]. The annotator describes the semantics of temporal locating adverbials found in texts as a succession of operators interacting with a temporal basis (date or event). Post-processing then associates calendar values to the adverbials using a formal algorithm described in [20]. This algorithm considers each operator that interacts with the temporal basis as a transformation on a calendar interval.

The normalization of deictic references is performed by a simple rule-based heuristic that only takes into account references containing calendar lexicon (day, month, etc.). It does not take into account the fact that a deictic reference may appear in a reported speech, nor verb tense, nor the possible ambiguity between an anaphoric or deictic interpretation (e.g. *at the beginning of the month* which can be either anaphoric or deictic). We assume that not normalizing all deictic references may not be a problem because of the volume of data.

Furthermore, as our evaluation protocol consists in comparing the outputs of our system with human chronologies edited by the AFP, we filter dates with a day granularity since in AFP hand-produced chronologies, we only find day granularity dates. If a date of smaller granularity is found (e.g. this morning, at 11am), we consider the day it is associated with. Though in these experiments the system filters dates on a specific granularity, it is nevertheless able to work on any granularity.

The output is a list of dates, ranked from most to least frequent with respect to the given topic. Each date is presented with a set of relevant sentences.

4.3 Salient Events Identification

The salient event extraction consists in extracting the most relevant sentence in each set (cluster) of sentences.

Semantic criterion for sentence filtering. The semantic analysis of temporal locating adverbials (see Section 3.2) allows us to distinguish a sentence that describes an event occurring inside the period of time denoted by an adverbial (In D_i) and a sentence that describes an event occurring after or before this period of time (Before D_i, After D_i or Since D_i, for instance). Therefore, the second criterion, a qualitative criterion, consists in selecting in each cluster C_{i1} the sentences that contain adverbials of the type In D_i. The resulting cluster C_{i2} is smaller than or equal to C_{i1}.

In the corpus used in our experiments, which mainly concerns present events (hence the importance of deictic references), this criterion is not very discriminating, since less than 3% of the adverbials are not of the type In D_i. Nevertheless, it avoids noise in the results and it is a useful criterion to consider in our global approach.

Discovering patterns. In order to discover relevant patterns, the process consists of two classical steps: a training step and a validation step. The training step aims at tuning the parameters of the sequential data mining. These parameters are as follows: the support, the min length, and the gap.

We used the Egyptian corpus as a training corpus (see section 4.1). A set of sentences

(filtered with a semantic criterion as explained previously) is associated to each salient date. There are 12 dates and therefore 12 sets of sentences (12 corpora) containing between 55 and 127 sentences. Each corpus is used as a sequential database (a sequence represents a sentence). In our experiments, supports varied from 10 to 4 (beyond 10 we get no pattern, and below 4 we get some thousand, or million, of patterns), the min length from 5 to 8 (this range was chosen for the same reasons as the support range) and two gaps were tested, [0,0] (corresponding to contiguous words) and [0,1] (a gap of one word is allowed). For each experiment, we tested the resulting patterns to extract the event judged salient in the newswires using similarity (see the next sub-section). In this framework, we noted that the best configuration is when the gap is [0,1], min length is 5 and the number of patterns is about 10 or 20.

For the validation step, we kept these values for the gap and the min length. We ran an automatic iterative pattern mining, varying the support, until the number of patterns came close to about 10.

Extracting the best sentence. In order to extract the sentence that best represents a salient event, we evaluate the relevance of each sentence in a cluster associated to a date, measuring its similarity to every extracted pattern, using the Jaccard measure. The more similar it is to the patterns, the higher its score is.

4.4 Evaluation

We evaluate two aspects of the system: its ability to provide relevant salient dates and its ability to provide relevant salient events, considering a given topic and a given period of time. In this sense, we compare the chronologies produced by the system to those of the AFP. The evaluation is performed on three of the sub-corpora (Irak, Sudan and PSA chronologies); the first one was kept for training the system (Egypt chronology).

To evaluate our system's outputs, we compare its results with hand-produced AFP chronologies. This is why we set the K parameter so that it corresponds to the number of events described in a given chronology. We therefore obtain for each chronology a set of K sentences representing K most salient events for a given topic in the period of time covered by the initial chronology.

Regarding the salient dates, the evaluation shows that the system has a recall rate of 69.2%: this rate measures the overlap between the set of salient dates provided by the system and the dates present in the AFP chronologies.

Regarding the salient events, the evaluation consists in comparing the system's output on this task –a sentence that describes an event– with the sentence associated to a given date in the hand-produced AFP chronologies. The evaluation shows that 65.3% of the sentences extracted by the system well reflects the event described in the sentences of the AFP chronology. Each sentence provided by the system was reviewed by two evaluators (4 evaluators were involved in the evaluation). Since this task can be subjective, we measured agreement between each pair of evaluators, which appears to be a good agreement (Kappa = 0.73).

Therefore if a comparison made by human evaluators between the manually- and system-generated timelines showed that although manually-generated timelines are on average preferable, our system is efficient and gives promising results which are close to manually generated timelines.

5 Conclusion

We have described a system for extracting salient events (represented by sentential textual units) from a large corpus to be placed along a timeline given a query (i.e. a time span and a topic) from the user. This allows users to rapidly obtain overviews of events related to their queries.

Unlike previous work both on summarization or event extraction using temporal criteria and on timelines building, we take into account not only the values of dates but also the textual units to which they belong, namely, temporal locating adverbials. We have shown that relying on this linguistic principle is crucial because it makes possible to use data mining techniques combined with symbolic NLP approach (see the beginning of Section 3.1 for a fuller explanation).

Currently, we are working on refining our annotator of temporal deictics by taking into

account their possible presence in direct reported speech. We are also working on the possibility given to the user or to the system (depending on the input corpus) to choose the granularity of events (for the moment, Day granule is the default granularity in our system) and the number K of events (for the moment, K is the number of events as it appears in hand-produced AFP chronologies).

Moreover, for the event extraction, we observed that discovered patterns contain very relevant information regarding the salient event. We are now investigating the process for selecting the most two or three informative patterns (instead of a sentence from the text) to be presented as the salient event.

Acknowledgements

This work has been partially funded by ANR Chronolines and Ecos-Sud 28 80.

References

1. **McKeown, K.R., Hatzivassiloglou, V., Barzilay, R., Schiffman, B., Evans D., & Teufel, S. (2001).** Columbia multi-document summarization: approach and evaluation. *Proceedings of the Document Understanding Conference (DUC01),* New Orleans, Louisiana, USA.
2. **Barzilay, R., Elhadad, N., & McKeown, K.R. (2002).** Inferring Strategies for Sentence Ordering in Multidocument News Summarization. *Journal of Artificial Intelligence Research*, 17(1), 35–55.
3. **Mani, I. & Wilson, G. (2000).** Robust temporal processing of news. *38th Annual Meeting on Association for Computational Linguistics (ACL'00)*, Hong Kong, China, 69–76.
4. **Li, Z., Wang, B., Li, M., & Ma W.Y. (2005).** A Probabilistic Model for Restrospective News Event Detection. *28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05), Salvador, Brazil, 106–113*.
5. **Smith, D.A. (2002).** Detecting and Browsing Events in Unstructured Text. *25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '02),* Tampere, Finland, 73–80.
6. **Swan, R. & Allan, J. (2000).** Automatic Generation of Overview Timelines. *23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '00)*, Athens, Greece, 49–56.
7. **Allan, J., Gupta, R., & Khandelwal, V. (2001).** Temporal summaries of new topics. *24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01)*, New Orleans, Louisiana, USA, 10–18.
8. **Chieu, H.L. & Lee, Y.K. (2004).** Query based event extraction along a timeline. *27th Annual International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR '04)*, Sheffield, UK, 425–432.
9. **Yan, R., Kong, L., Huang, C., Wan, X., Li, X., & Zhang, Y. (2011).** Timeline generation through evolutionary trans-temporal summarization. *Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*, Edinburgh, UK, 433–443.
10. **Kessler, R., Tannier, X., Hagège, C., Moriceau, V., & Bittar, A. (2012).** Finding Salient Dates for Building Thematic Timelines: Long Paper. *50th Annual Meeting of the Association for Computational Linguistics (ACL '12),* Jeju Island, Korea, 1, 730–739.
11. **Cellier, P., Charnois, T., Plantevit, M., & Crémilleux, B. (2010).** Recursive Sequence Mining to Discover Named Entity Relations. *Advances in Intelligent Data Analysis IX*, *Lecture Notes in Computer Science,* 6065, 30–41.
12. **Pustejovsky, J., Castaño, J., Ingria, R., Sauri, R., Gaizauskas, R., Setzer, A., & Katz, G. (2003).** TimeML: Robust Specification of Event and Temporal Expressions in Text. *Fifth International Workshop on Computational Semantics (IWCS-5),* Tilburg, Netherlands.
13. **Battistelli, D., Couto, J., Minel, J.L., & Schwer, S.R. (2008).** Representing and Visualizing calendar expressions in texts. *2008 Conference on Semantics in Text Processing (STEP'08)*, Venice, Italy, 365–373.
14. **Teissèdre, C., Battistelli, D., & Minel, J.L. (2010).** Resources for Calendar Expressions Semantic Tagging and Temporal Navigation through Texts. *7th International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta, 3572–3577.
15. **Agrawal, R. & Srikant, R. (1995).** Mining sequential patterns. *Eleventh International Conference on Data Engineering (ICDE '95), Taipei, Taiwan, 3–14*.

16. **Srikant, R. & Agrawal, R. (1996).** Mining sequential patterns: Generalizations and performance improvements. *Advances in Database Technolohy–EDBT'96, Lecture Notes in Computer Science*, 1057, 1–17.
17. **Yan, X., Han, J., & Afshar, R. (2003).** CloSpan: Mining closed sequential patterns in large databases. *Third SIAM International Conference on Data Mining*, San Francisco, California.
18. **Zaki, M.J. (2001).** SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning,* 42(1-2), 31–60.
19. **Dong, G. & Pei, J. (2007).** Sequence Data Mining. New York: Springer.
20. **Battistelli, D., Cori, M., Minel, J.L., & Teissèdre, C. (2011).** Semantics of Calendar Adverbials for Information Retrieval. *Foundations of Intelligent Systems*, *Lecture Notes in Computer Science,* 6804, 622–631.

Delphine Battistelli is an assistant professor at University of Paris Sorbonne. Her main research field is temporality and modality analysis in texts. She presently conducts several projects in this field which involve NLP and data mining techniques.

Thierry Charnois is an assistant professor at the GREYC Laboratory (CNRSUMR 6072), University of Caen, France. He holds a Ph.D. in Computer Science (1999) from LIPN Laboratory, University of Paris 13, and a Habilitation degree in 2011 from the University of Caen. His research interests include natural language processing, knowledge discovery, information extraction from texts, and their applications, notably in bioinformatics, digital humanities or opinion analysis.

Jean-Luc Minel is a full time professor at University of Paris Ouest Nanterre La Défense and the head of MoDyco, a CNRS laboratory. He is presently involved in several national and international projects which combine different linguistic analysis and data mining techniques.

Charles Teissèdre is a temporary research and teaching assistant and member of the STIH laboratory, at Paris-Sorbonne University, France. His research interests include natural language processing, information retrieval, information extraction from texts as well as knowledge representation and acquisition.

Article received on 05/12/2012; accepted on 17/01/2013.