Automatic Identification of Temporal Information in Tourism Web Pages

Stéphanie Weiser*, Philippe Laublet, Jean-Luc Minel***
* MoDyCo, UMR 7114, CNRS
200 avenue de la République, 92001 Nanterre
** LaLIC, Université Paris-Sorbonne
Maison de la recherche, 28 rue Serpente 75006 Paris
E-mail: steph.weiser@gmail.com, Philippe.Laublet@paris-sorbonne.fr, jminel@u-paris10.fr

Abstract

This paper presents our work on the detection of temporal information in web pages. The pages examined within the scope of this study were taken from the tourism sector and the temporal information in question is thus particular to this area. The differences that exist between extraction from plain textual data and extraction from the web are brought to light. These differences mainly concern the spatial arrangement of the text, the use of punctuation and the respect of traditional syntactic rules. The temporal expressions to be extracted are classified into two kinds: temporal information that concerns one particular event and repetitive temporal information. We adopt a symbolic approach relying on patterns and rules for the detection, extraction and annotation of temporal expressions; our method is based on the use of transducers. First evaluations have shown promising results. Since the visual structure of a web page is very important and often informs the user before he has even read the text, a semiotic study is also presented in this paper.

1. Introduction

With the methods of the Semantic Web, portal applications can be created, relying on ontologies. For these applications and many service applications, temporal information is often essential. For example, a tourism web portal would need information about the type of tourism object and its location in time and space. In addition, the extracted information must be stored in the knowledge base according to the ontology used by the application.

In this paper we will focus on temporal information in tourism web pages. The temporal information has to be detected, extracted and annotated. The annotation format will probably rely on existing XML tools (Stern 2007). To perform these tasks, we encountered three main kinds of difficulties. First, we have to deal with complex, imprecise temporal information. Of course, single dates are easy to process but more complex expressions, such as periods or repetitive information (e.g. *from March to July, open every day except Tuesday*), must be treated as well. Second, after being extracted, the information needs to be linked to the proper tourism object. If the web page concerns only one object, this is straightforward, but some web pages concern many objects and an analysis is therefore necessary to decide how to link each piece of information to its object. Third, the web pages we deal with are all of the same type: tourism web pages. However, they vary a lot as they are made by different people, have different forms and concern different types of tourism objects. We will try to show that a semiotic study of some pages is necessary to take into account some of their specificities.

The work presented in this paper is situated within the framework of the EIFFEL project. Its main objective is to create a portal in the area of tourism with different functionalities. This portal, for use by the local tourism sector, will include a specialised search engine. It will allow users to find and collect precise and essential information in context. It will also help the territory as a French region to promote its services. This is a wide-ranging project (Noël et al. 2008) which is based on web semantic technologies, knowledge representation and linguistic methods and expertise. It includes automatic identification, selection and extraction of various items from the Web according to existing ontologies.

This project involves mainly two companies – Mondeca and Antidot – and three laboratories – LIRMM (CNRS), INRIA-Rocquencourt and MoDyCo (CNRS).

Our corpus, meaning that the pages have been collected for this particular study, in a precise context, is composed of more than 5000 tourism web pages in French. These pages were collected by automatic crawling of websites. They were collected by the company Antidot and transformed into XML files in order to make them more suitable for automatic processing.

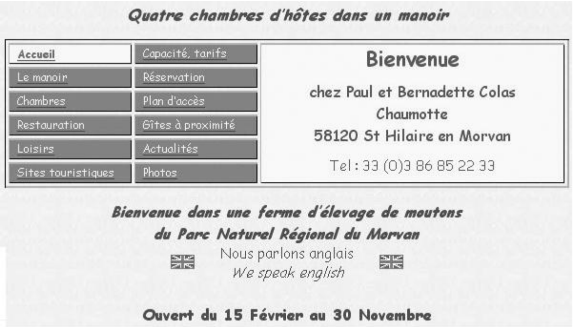
2. Temporal Information in Web Pages

The detection and extraction of temporal information is a subject that has already been widely studied (Battistelli et al. 2006), but mostly in the context of "printed textual objects" and not in texts specifically written for web pages. One of the aims of this paper is to show that studying the same type of temporal information is very different when it takes place in a web context.

In tourist guide books, for example, the information is a lot more structured and standardized than on our web sites, which are extremely diverse and do not follow a common page model. Every entry of a guide concerning a particular tourism object is often structured in the same way and contains the same type of information. Linguistic markers allow us to locate each piece of information. Punctuation is very rigorous and facilitates interpretation, especially automatic interpretation. On the other hand, on web pages, punctuation – if there is any – is rarely correctly used. More often, white spaces or new lines are used as separators.

Even if the analysis of a text is not easy, a lot of syntactic tools already exist to perform this task automatically. For example, it is possible to build the syntactic tree of a given sentence, find the verb, etc. However, some complicated tasks, such as anaphora resolution, can make the analysis of texts a lot more complex. What we are interested in here is that the difficulties that are encountered for web pages are very different from those that text analysis presents.

For example, in a tourism guide book[1], each restaurant entry is structured in the same way: name, address, phone number, metro station and opening details, e.g. *ouvert tous les jours* (*open every day*). Each item can easily be automatically processed. The problem is that, on a web page, the same kind of information would more likely look like the web page presented in figure 1.



Quatre chambres d'hôtes dans un manoir

Bienvenue

chez Paul et Bernadette Colas
Chaumotte
58120 St Hilaire en Morvan

Tel : 33 (0)3 86 85 22 33

Bienvenue dans une ferme d'élevage de moutons
du Parc Naturel Régional du Morvan

Nous parlons anglais
We speak english

Ouvert du 15 Février au 30 Novembre

Figure 1: Example

For now, there is no tool able to interpret that *Chez Paul et Bernadette Colas* is the name of a Bed & Breakfast or that *Ouvert du 15 février au 30* (*open from February 15th to 30th*) is its opening information. This is why the extraction of temporal information from texts can not be directly applied for Web pages: there is no standardized "syntax" for web pages.

Our work is close to that of (Tenier et al. 2006). The object of their work is to extract, from web pages, information about contributors to scientific events. Our approach is more general since the web pages analyzed by these authors follow spatial arrangement rules while spatial arrangement on our pages is completely free. Our work is also close to that of (Bry et al. 2003) who aim at integrating temporal and spatial reasoning into XML query and transformation operations. Their work has lead to WebCal, a program which provides calendrical calculations for web services.

3. Tool and Method for the Temporal Expression Detection

We have a symbolic approach relying on patterns and rules for the detection, extraction and annotation of temporal expressions (in French). We do not use any machine learning techniques. Our method relies on the use of transducers, created using the graphical interface of the Unitex corpus processor[2]. Unitex provides an intuitive and user-friendly method for creating transducers, which tag the output (as opposed to regular expressions which do not modify the output). A wide variety of patterns can be recognized and tagged by our transducers, from very simple expressions to very complex ones. On one hand, generalizations can be made (such as using the symbol <NB> to tag all numbers). On the other hand, the transducers can be very detailed and precise for particular cases, for example, to recognize expressions such as *à l'exception du mois de mai – except in May*.

Here are a few examples of the expressions to be detected:

(1) Le 21 avril
April, the 21st
(2) Du 10 au 21 mars
From the 10th to the 21st of March
(3) Du lundi au vendredi, 9h – 11h
From Monday to Friday, 9am-11am
(4) Inauguration du musée le 7 juillet 2006
Inauguration of the museum on the 7th of July 2006
(5) En Juillet et Août ouvert tous les jours. Hors cette période, fermeture le mercredi soir, le jeudi toute la journée et le dimanche soir. Horaires d'ouverture: de 12h00 à 14h00 de 19h00 à 22h00.
Open every day in July and August. Outside this period, closed on Wednesday night, all day Thursday and on Sunday night. Opening times: from noon to 2pm from 7pm to 10pm.

3.1. Classification

As shown in these examples, the type of temporal information to be identified varies. Two main types can be highlighted: temporal information that concerns one particular event and repetitive temporal information. For the first type there are dates (*concert on October 1st*), periods (*festival from May to June*), and times (*the concert starts at 8:00*). For the second there are times (*the museum opens at 10am*), periods (*the restaurant is open from Monday to Saturday*) and exceptions (*the camping ground is open all year long except in January*).More complex examples, such as *from March*

[1] *Le guide du routard – Paris balades* 2007 p.231 : « Le Saint-Amour 2, av.Gambetta, 75020. 01-47-97-20-15. Métro Père-Lachaise. Ouvert tous les jours » (*Open every day*)

[2] Unitex: http://www-igm.univ-mlv.fr/~unitex

to July, open every day except Tuesday, can also be included in this classification.

In addition, not every occurrence of temporal information needs to be detected, since they do not necessarily concern a tourism object. For example, a single date, which is not introduced by a linguistic marker such as *heure d'ouverture* (*opening time*) probably does not concern a tourism object. Single dates like these often concern the web page itself as in *Dernière modification : 25/10/2006* (*last modification : 25/10/2006*). Therefore, we have been careful not to detect these dates.

One problem that has been encountered is that some lexical marks are ambiguous. For example is *Friday and Sunday night* to be interpreted as meaning Friday night and Saturday night, or rather Friday all day and Saturday night? The context is therefore necessary to remove the ambiguity. For this example, if the object is a concert, the first interpretation will probably be the right one.

3.2. Technical Information

The tool Unitex, which uses dictionaries based on the tables of the LADL[3], allows the user to process corpora on the lexical, syntactic and morphological levels. It allows for the identification and tagging of structures which correspond to regular expressions, represented by finite state graphs. This graphical representation makes using Unitex more intuitive than the use of long regular expressions. We built our transducers using this tool, following the linguistic study that listed the patterns of temporal expressions that could be found on web pages. The set of transducers includes one main transducer and 23 secondary graphs. These transducers include 88 lexical markers (like *ouverture – opening, mars – March, inauguration*) and 22 grammatical markers (like *sur – on, le – the, dans – in*).

4. Evaluations

4.1. First Evaluation

The transducers allowing the detection of temporal information were manually created from a study of a set of pages. In order to evaluate and optimise them, we tested them on sets of 20 web pages, randomly manually selected from our corpus. This operation was performed five times and the transducers have thus been tested on 100 different pages.

The results of this experiment are presented in the diagrams below. Recall and precision were calculated for each set of 20 pages (figure 2 and figure 3).

For the whole corpus of 100 pages, we obtain a recall rate of 92% and a precision of 76%.

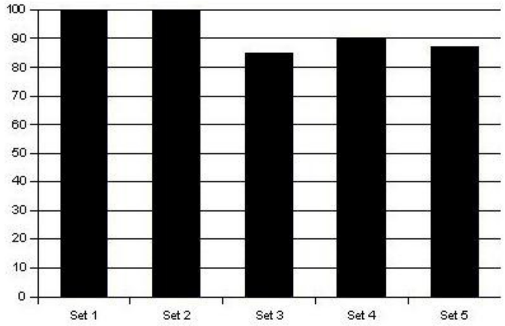


Figure 2: Recall for the first evaluation

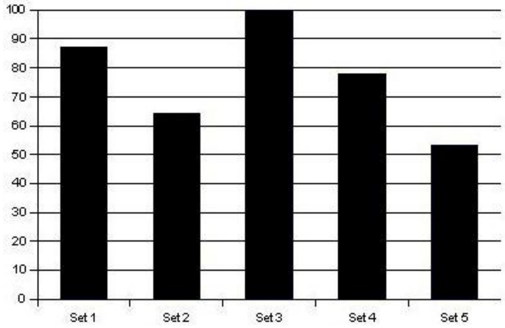


Figure 3: Precision for the first evaluation

The important variation in the results is due to the fact that there are not a lot of temporal expressions to detect in the pages (7 for sets 1 and 3, 9 for set 2, 20 for set 4 and 8 for set 5). The precision for set 5 is especially low because one page contains a forum with 4 temporal expressions that are detected but do not concern a tourism object.

4.2. Second Evaluation

A different procedure was used for the second evaluation. We deliberately manually selected 25 pages containing temporal information. Some of these pages were randomly opened and selected when they were found to contain temporal information; for the others our transducers were applied to more than 200 pages and those with results were selected.

For these 25 pages, the recall is 39,1% and the precision is 46,5%. Here are a few numbers to explain these figures. These pages contain 69 expressions to detect. The transducers match and tag 58 expressions: 27 are correctly detected, 3 are falsely detected (they should not be detected because they do not correspond to temporal tourism information), 28 expressions are incompletely detected and 25 expressions are missed.

We are faced with two kinds of incomplete expressions. In the first kind a part of the expression is not matched,

[3] Laboratoire d'Automatique Documentaire et Linguistique – the tables were created by Maurice Gross and consist in a classification of lexical units according to syntactic and distributional criteria.

for example in *toute la journée* de 8h00 à 19h00 (*all day long* from 8:00am to 7:00pm), only the underlined part is detected. In the second kind of incomplete expressions, each pattern to be matched is detected in two or more matches, for example ouvert tous les jours en juillet et août (open every day in July and August) is detected in two matches. 16 of the 28 incomplete expressions produced by our system correspond in fact to 5 correct expressions.

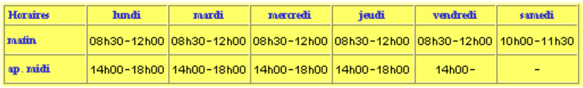
The high number of missed expressions is due to the fact that isolated dates without context are voluntarily ignored in order to avoid incorrect detections (like in *dernière modification : 25 avril 2006 – last modification: April the 25th 2006*). Some pages presenting an event schedule contain more than 10 dates without context that are therefore not detected. 22 out of the 25 missed expressions are concerned by this problem, which will be solved in the near future.

5. The Semiotic Marks

For now, only textual information is taken into account, but a web page presents information in many other ways. The visual structure of a web page often informs the user before he has even read the text. For example, the structure could allow him to understand that the page concern one single object and not a collection of objects. In the same way, the formatting style (font, size, colour) can be very meaningful since it can often be assumed that two items written in the same font and colour are linked.

One difficulty lies in automatically interpreting these semiotic marks. Many things are understandable at first sight for a human being but are hard to specify for automatic processing.

On web pages, tables are commonly used to present information. They need a special treatment since their interpretation is not straightforward.



Horaires	lundi	mardi	mercredi	jeudi	vendredi	samedi
matin	08h30-12h00	08h30-12h00	08h30-12h00	08h30-12h00	08h30-12h00	10h00-11h30
ap. midi	14h00-18h00	14h00-18h00	14h00-18h00	14h00-18h00	14h00-	-

Figure 4: Opening times table

The figure 4 is an example of a table of opening times found on a web page. The first row contains the names of the days, the second row contains the opening times in the mornings and the last row contains the opening times in the afternoons.

The information could have been presented in a vertical table as the one shown in figure 5:

Horaires	Matin	Ap-midi
Lundi	8h30-12h00	14h00-18h00
Mardi	8h30-12h00	14h00-18h00
Mercredi	8h30-12h00	14h00-18h00
Jeudi	8h30-12h00	14h00-18h00
Vendredi	8h30-12h00	14h00-
Samedi	10h00-11h30	-

Figure 5: vertical table of opening times

A human user would understand both of these tables without the slightest problem. For an automatic treatment, however, it is more complicated. If we do not look at the structure, but only at the text, the first table would then contain "Horaires lundi mardi mercredi jeudi vendredi samedi matin 08h30 - 12h00 08h30 - 12h00 08h30 - 12h00 08h30 - 12h00 08h30 - 12h00 10h00 - 11h30 ap. midi 14h00 - 18h00 14h00 - 18h00 14h00 - 18h00 14h00 - 18h00 14h00 - - ": a text flow that is incomprehensible. The second table would contain "Horaires Matin Ap-midi Lundi 8h30-12h00 14h00-18h00 Mardi 8h30-12h00 14h00-18h00 Mercredi 8h30-12h00 14h00-18h00 Jeudi 8h30-12h00 14h00-18h00 Vendredi 8h30-12h00 14h00 – Samedi 10h00-11h30 - " it is clearer but it is very different from the text of the first one, for exactly the same meaning.

Technical tools that allow a specific markup for headers in the table exist. But the web pages of our corpus are mostly built by professionals from the tourism sector and not by professional web designers: they do not use these technical tools. Consequently, only graphical elements and lexical content allow interpretation of these kinds of tables.

In order to take into account these semiotic marks in automatic processing, the structure of the pages must be studied. An analysis of the XML tree is therefore necessary (Tenier et al. 2006).

6. Conclusion

In this paper we have presented our work on the detection of temporal information in web pages. We have focused on the differences that exist between extraction from plain textual data and extraction from the web. In addition, we introduced our main linguistic resource, our transducers, which could be used for many different NLP applications. We also presented the results obtained in applying these transducers to our corpus.

The transducers that have been presented will be improved. They will be enriched with information concerning the type of tourism object, in order to determine the links between the temporal information detected and what they refer to.

The work on semiotic marks will be continued in order to implement the study of the structure of the pages, analysing XML trees.

7. Acknowledgements

The project EIFFEL (ANR-05-RNTL-007) is funded with an ANR grant.

8. References

Battistelli, D., Minel, J.-L., Schwer, S. (2006). Représentation des expressions calendaires dans les textes : une application à la lecture assistée de biographies, *Traitement Automatique des Langues*, 47, 3, pp.1--26.

Bry, F. Lorenz, B. Ohlbach, H. J. Spranger, S. (2003). On Reasoning on Time and Location on the Web, *Lecture Notes in Computer Science*, Springer-Verlag, Germany, pp. 69--83.

Noël, L., Carloni, O., Moreau, N., Weiser, S. (2008). Designing a Knowledge-Based Tourism Information System, *Int. J. of Digital Culture and Electronic Tourism*, Special Issue on National Tourism Organisations and Exploitation of Information Technologies, to be published.

Stern, R.-D. (2007). Expression linguistique du temps et représentation ontologique : OWL-Time et étude des adverbiaux temporels, Mémoire de Master IILGI, Université de Paris-Sorbonne.

Tenier, S., Toussaint, Y., Napoli, A. et Polanco, X. (2006). Instantiation of relations for semantic annotation, *In the 2006 IEEE/WIC/ACM International Conference on Web Intelligence - WI 2006*, pp. 463--472