French Resources for Extraction and Normalization of Temporal Expressions with HeidelTime

Véronique Moriceau, Xavier Tannier

LIMSI-CNRS, Univ. Paris-Sud
91403 Orsay, France
{moriceau, xtannier}@limsi.fr

Abstract

In this paper, we describe the development of French resources for the extraction and normalization of temporal expressions with HeidelTime, a open-source multilingual, cross-domain temporal tagger. HeidelTime extracts temporal expressions from documents and normalizes them according to the TIMEX3 annotation standard. Several types of temporal expressions are extracted: dates, times, durations and temporal sets. French resources have been evaluated in two different ways: on the French TimeBank corpus, a corpus of newspaper articles in French annotated according to the ISO-TimeML standard, and on a user application for automatic building of event timelines. Results on the French TimeBank are quite satisfaying as they are comparable to those obtained by HeidelTime in English and Spanish on newswire articles. Concerning the user application, we used two temporal taggers for the preprocessing of the corpus in order to compare their performance and results show that the performances of our application on French documents are better with HeidelTime. The French resources and evaluation scripts are publicly available with HeidelTime.

Keywords: Temporal expressions, normalization, French resources

1. Introduction

The analysis of temporal information is often an essential component in text understanding and is useful in a wide range of information retrieval applications (Alonso et al., 2007; Alonso, 2008; Kanhabua, 2009; Mestl et al., 2009). The task of temporal annotation consists in extracting and normalizing temporal expressions. Normalization is the operation of turning a temporal expression into a formated, fully specified representation (this includes finding the absolute value of relative dates). The TempEval challenges, for example, focus on the evaluation of temporal information processing using the ISO-TimeML language (Pustejovsky et al., 2010), a specification language for manual annotation of temporal information in texts.

HeidelTime is a multilingual, cross-domain temporal tagger which extracts temporal expressions from documents and normalizes them according to the TIMEX3 annotation standard (Strötgen and Gertz, 2013). It is an open-source tagger which achieved the best results for the extraction and normalization of temporal expressions for English documents in the context of the TempEval-2 and TempEval-3 challenges (Verhagen et al., 2010; Uzzaman et al., 2012). HeidelTime processes documents in English, German, Dutch, Vietnamese, Arabic, Spanish, Italian and we developed French resources that are now publicly available within the official HeidelTime distribution[1].

The paper is structured as follows: we first present HeidelTime. Section 3 describes the development of French resources. Finally, we present and discuss the evaluation results in Section 4. This evaluation is twofold: first, HeidelTime with French resources has been evaluated on the French TimeBank corpus and then on a user application which automatically builds event timelines.

[1] http://code.google.com/p/heideltime/

2. Presentation of HeidelTime

HeidelTime is a multilingual, cross-domain temporal tagger. It is a rule-based system with a separation between language-dependent resources and generic Java code. Resources consist in extraction rules (regular expression patterns) and lexicons for normalization (for example, weekdays, months, etc). Absolute temporal expressions (*July 26th, 2013* ; *07-26-2013*) are extracted and normalized by the extraction rules. Relative temporal expressions (*yesterday, July*) are extracted by the rules but are left underspecified at this step.

Then, a normalization (Java code) is applied according to the type of documents (news, scientific, etc.) and to the tense of the verb used in the sentence. As a reference time, normalization can use the document creation time (DCT) or the previously mentioned date. HeidelTime requires sentence, token, and part-of-speech information. For English, the TreeTagger is used. Figure 1 shows an example of normalization for an absolute date and a DCT relative expression. Attribute *type* is the type of the extracted temporal expression and attribute *value* is its normalization.

DCT: 2009-12-22

Spaniards often choose lottery numbers matching significant dates. One of the most requested ticket numbers <TIMEX3 *tid="t1" type="DATE" value="2009"*>this year</TIMEX3> was 25609, which corresponds to <TIMEX3 *tid="t2" type="DATE" value="2009-06-25"*> June 25, 2009</TIMEX3>, the day pop star Michael Jackson died.

Figure 1: Example of HeidelTime output.

3. Development of French Resources

Resources are composed of 3 types of files read by HeidelTime's resource interpreter and which have to follow HeidelTime's rule syntax: patterns, normalizations and rules. The pattern files contain words and phrases used to express temporal expressions (months, days, etc.). The normalization files contain normalization information about the patterns (for example, the normalized value of *February* is *02*). Finally, the rule files contain rules for date (100 rules), time (20 rules), duration (25 rules), and set expressions (12 rules). All rules have:

- an extraction part which defines the expressions that have to be matched in a document, using the pattern resources,
- a normalization part which normalizes the extracted expression using the normalization resources.

For example, the following rule is used to extract and normalize the temporal expressions *jeudi 4 octobre 2012* or *le lundi 23 sept. 2013*:

```
RULENAME="date_r1",
EXTRACTION="([Ll]e )g1%reWeekdayg2
%reDayNumberg3
(%reMonthLongg5|%reMonthShortg6)g4
%reYear4Digitg7",
NORM_VALUE="group(7)-%normMonth(group(4))-
%normDay(group(3))"
```

where:
- `group(7)` is the string extracted with the pattern `%reYear4Digit` (*i.e.* 2012 or 2013),
- `%normMonth(group(4))` is the normalization of *octobre* extracted with the pattern `%reMonthLong` (*i.e. 10*) or *sept.* extracted with the pattern `%reMonthShort` (*i.e. 09*),
- `%normDay(group(3))` is the normalization of *4* (*i.e. 04*) or *23* (*i.e. 23*) extracted with the pattern `%reDayNumber`.

The normalized values are then *2012-10-04* and *2013-09-23*.

To develop the French pattern and normalization resources, we translated the English and Spanish resources and adapted them to the French patterns. To develop the French rules for dates, times, durations and sets, we used a corpus composed of 350 news articles from Agence France Presse (AFP). Note that during the process of rule development, an important point that has to be taken into account is that French is an inflected language (nouns, adjectives, determiners, verbs).

HeidelTime uses the tense of the sentence verb to determine if a temporal expression refers to a past or future date w.r.t. the DCT. This information is given by the French TreeTagger which is used for preprocessing the French documents (sentence, token and part-of-speech annotation). For example, the following rule is used to extract and normalize the temporal expressions *mars* in *Il est parti en mars* (*He left in March*) or in *Il reviendra en mars* (*He will come back in March*) :

```
RULENAME="date_r7",
EXTRACTION=
"(%reMonthLongg1|%reMonthShortg2)g3",
NORM_VALUE="UNDEF-year-%normMonth(group(1))"
```

where:
- `%normMonth(group(1))` is the normalization of *mars* (*i.e. 03*) extracted with the pattern `%reMonthLong`,
- `UNDEF-year` is an undefined year that is then calculated with the DCT and the tense of the verb. The DCT of both documents is *2009-09*. In *Il est parti en mars* (*He left in March*), a past tense verb is identified so the normalized value is *2009-03-XX* whereas in *Il reviendra en mars* (*He will come back in March*), a future tense verb is identified so the normalized value is *2010-03-XX*.

Figure 2 shows some examples of extraction and normalization for each temporal type.

DCT: 1999-07-07
La France a vu sa population augmenter de plus de 2 millions d'habitants en <TIMEX3 *tid="t1" type="DURATION" value="PY9"*>9 ans</TIMEX3>. A l'aube de l'an <TIMEX3 *tid="t2" type="DATE" value="2000"*>2000</TIMEX3>, sa population s'établissait <TIMEX3 *tid="t3" type="DATE" value="1999-03-08"*>le 8 mars dernier</TIMEX3> à 60082000 habitants.
*(France saw its population increase by more than 2 million people in **9 years**. At the dawn of **2000**, the population stood **on March, 8** at 60 082 000 inhabitants.)*

DCT: 1999-05-18
<TIMEX3 *tid="t1" type="TIME" value="1999-05-23TEV"*>Dimanche soir</TIMEX3>, à partir de <TIMEX3 *tid="t2" type="TIME" value="1999-05-23T22:00"*>22 h</TIMEX3>, le comité des fêtes vous invite également au bal.
*(**Sunday evening**, from **22 pm**, the festival committee also invites you to a ball.)*

DCT: 2002-02-09
Quelque 9 millions de personnes visitent <TIMEX3 *tid="t1" type="SET" value="P1Y"*>chaque année</TIMEX3> les parcs nationaux dans l'Utah.
*(9 million people **annually** visit the national parks in Utah.)*

Figure 2: Examples of HeidelTime output on French documents.

4. Evaluation

The French resources have been evaluated in two different ways: on the French TimeBank corpus and on a user application for automatic building of event timelines. We present in this section the results obtained for both evaluations.

4.1. Evaluation Results on the French TimeBank

We first evaluated HeidelTime with the French resources on the French TimeBank corpus[2] (Bittar et al., 2011),

[2] `https://gforge.inria.fr/projects/fr-timebank/`

composed of 108 newspaper articles in French annotated according to the ISO-TimeML standard. In this corpus, there are 425 temporal expressions in texts, among which 227 dates, 130 time expressions, 52 duration expressions and 16 temporal sets.

To evaluate the extraction performance, we used the measures used in TempEval challenges: precision, recall and F1-score for strict and relaxed matching. We also computed the F1 measures on the two most important TIMEX attributes *value* and *type*. The *value F1* score captures the performance of the system to extract a temporal expression with a correct normalization. The *type F1* is the performance of the system to extract a temporal expression with a correct type (*i.e.* date, time, duration, set). Evaluation scripts that we developed are also publicly available[3].

4.1.1. Results

Table 1 presents the results on the French TimeBank:

	Precision	Recall	F1
Strict match	0.86	0.84	0.85
Relaxed match	0.92	0.89	0.91
Value F1	0.74		
Type F1	0.83		

Table 1: HeidelTime's results with French resources on French TimeBank.

These results are quite satisfaying as they are comparable to those obtained by HeidelTime in English and Spanish on newswire articles (Strötgen et al., 2013). Indeed, on the TempEval 3 English corpus, F-score for strict matching is 0.81, attribute *value* F1 is 0.77 and attribute *type* F1 is 0.82. On the TempEval 3 Spanish corpus, F-score for strict matching is 0.85, attribute *value* F1 is 0.85 and attribute *type* F1 is 0.87 (if values are similar, results are hardly comparable, since the corpora are different).

Table 2 presents the detailed results for each temporal type. Here, a matching is correct if there is a strict or relaxed matching and if the *type* attribute is correct. A *value* attribute is considered as correct only if the matching is correct.

	Correct Match		Correct Match & Correct Value		
Total ①	# ②	% w.r.t ①	#	% w.r.t ①	% w.r.t ②
DATE (227)	212	93.4 %	187	82.4 %	88.2 %
TIME (130)	84	64.6 %	62	47.7 %	73.8 %
DURATION (52)	40	76.9 %	40	76.9 %	100 %
TEMPORAL SET (16)	8	50 %	6	37.5 %	75 %

Table 2: Detailed results on French TimeBank.

[3] http://code.google.com/p/heideltime/wiki/ReproduceEvaluationResults

4.1.2. Error Analysis

As we can see from Table 2, most extraction errors are on TIME and SET expressions.

For time expressions, we noticed in the French TimeBank that adverbs *maintenant, aujourd'hui* and *désormais* (*now, today, henceforth*) are inconsistently annotated either as a TIME or a DATE expression and that their value is either PRESENT-REF or a normalized value. In our French resources, we considered that these adverbs are DATE expressions. We found 22 occurrences of mismatches due to this problem. Another error cause is when a date is associated with a time expression (for example, *the interview will be on <TIMEX3 type="DATE" value="2012-06-05">June, 5th</TIMEX3> at <TIMEX3 type="DATE" value="2012-06-05T17:00">5 pm</TIMEX3>*): in this case, when the date normalization is incorrect then the time normalization is also incorrect.

Concerning duration expressions, some have not been extracted mainly because we did not develop rules for very specific cases (8 occurrences): for example, time expressions expressed in minutes or seconds, durations like *half-century*, *quinquennium* or *greco-roman period*, etc. But we can note that when a duration expression is correctly extracted, it is always normalized correctly.

4.2. Evaluation on a User Application

We have also evaluated HeidelTime with the French resources on a user application that automatically builds event timelines from a search query.

4.2.1. User Application: Event Timelines

We developed an approach for detecting salient (important) dates in texts in order to automatically build event timelines from a search query (Kessler et al., 2012b). In order to extract salient dates that warrant inclusion in an event timeline, a newswire article corpus is first pre-processed and temporal expressions are normalized. Then, the corpus is indexed by the Lucene[4] search engine. Given a query, a number of documents are retrieved by Lucene. Dates are extracted from documents and ranked in order to show the most important ones to the user together with the sentences that contain them.

4.2.2. Document Collection

So far, our system used the linguistic analyzer XIP (Aït-Mokhtar et al., 2002) which performs a deep syntactic analysis, named entity recognition and extraction and normalization of temporal expressions for English and French. We wanted to evaluate the performance of our system with another free temporal tagger. Thus we used two temporal taggers for the preprocessing of the corpus, HeidelTime and XIP, in order to compare their performance for the user application.

We used a corpus of newswire texts provided by the AFP French news agency. The French AFP corpus is composed of 1 million texts that span the 2004-2011 period (499 documents/day in average and 390 millions words). Each document is an XML file containing a title, a date of creation

[4] http://lucene.apache.org/

(DCT), set of keywords, and textual content split into paragraphs. Note that absolute dates are quite infrequent in this corpus (about 7%).

4.2.3. Results

Processing runs were evaluated on 94 manually-written chronologies according to Mean Average Precision (MAP), which is a widely accepted metric for ranked lists. These chronologies (textual event timelines) are a specific type of articles written by AFP journalists in order to contextualize current events. These chronologies consist in a list of dates (typically between 10 and 20) associated with a text describing the related event(s). With the corpus processed by XIP, MAP is 0.60 (Kessler et al., 2012a) whereas it is 0.64 with the corpus processed by HeidelTime. This result shows that the performances of our application on French documents are better with HeidelTime than with XIP which achieved good results in the TempEval campaign (Verhagen et al., 2007; Hagège and Tannier, 2008).

The main cause for incorrect value normalization of underspecified expressions in AFP corpus is wrong tense identification: for normalization, HeidelTime considers the tense of the verb close to the temporal expression but this verb may not be the main verb of the sentence and be in a different tense. In the following example, the temporal expression *mercredi* (*Wednesday*) is not normalized correctly because the closest verb which is used for normalization (*devra (will have to)*), is in future tense whereas the main verb of the sentence in present tense should be considered for normalisation :

*François Hollande **assure** que le prochain président de la République **devra** "être l'inverse de Nicolas Sarkozy", dans un entretien à Libération **mercredi**.*

*(François Hollande **declares** that the next president **will have** to "be the opposite of Nicolas Sarkozy," in an interview with Libération on **Wednesday**)*

5. Conclusion

In this paper, we presented the French resources we developed for HeidelTime, an open-source multilingual, cross-domain temporal tagger that achieved the best results in the TempEval-3 challenge. Our French resources allow for the extraction and normalization of French temporal expressions with HeidelTime and are now publicy available. HeidelTime with French resources achieved good results in both evaluations that we performed. In future work, we intend to perform evaluations on other types of documents than news articles.

Acknowledgements

This work has been partially funded by French National Research Agency (ANR) under project Chronolines (ANR-10-CORD-010). We would like to thank the French News Agency (AFP) for providing us with the corpus. We also thank Jannik Strötgen and Julian Zell for their help with HeidelTime.

6. References

Salah Aït-Mokhtar, Jean-Pierre Chanod, and Claude Roux. 2002. Robustness beyond Shallowness: Incremental Deep Parsing. *Natural Language Engineering*, 8:121–144.

Omar Alonso, Ricardo Baeza-Yates, and Michael Gertz. 2007. Exploratory Search Using Timelines. In *SIGCHI 2007 Workshop on Exploratory Search and HCI Workshop*, San Jose, USA.

Omar Rogelio Alonso. 2008. *Temporal information retrieval*. Ph.D. thesis, University of California at Davis, Davis, CA, USA. Adviser-Gertz, Michael.

André Bittar, Pascal Amsili, Pascal Denis, and Laurence Danlos. 2011. French TimeBank : An ISO-TimeML Annotated Reference Corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, Portland, USA.

Caroline Hagège and Xavier Tannier. 2008. XTM: A Robust Temporal Text Processor. In *Computational Linguistics and Intelligent Text Processing, proceedings of 9th International Conference CICLing 2008*, volume LNCS 4919 of *Lecture Notes in Computer Science*, Haifa, Israel. Springer Verlag.

Nattiya Kanhabua. 2009. Exploiting temporal information in retrieval of archived documents. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (SIGIR 2009)*, Boston, USA.

Rémy Kessler, Xavier Tannier, Caroline Hagège, Véronique Moriceau, and André Bittar. 2012a. Extraction de dates saillantes pour la construction de chronologies thématiques. *Revue Traitement Automatique des Langues*, 53/2.

Rémy Kessler, Xavier Tannier, Caroline Hagège, Véronique Moriceau, and André Bittar. 2012b. Finding Salient Dates for Building Thematic Timelines. In *Proceedings of the 50th annual meeting of the Association for Computational Linguistics (ACL 2012)*, Jeju, Korea.

Thomas Mestl, Olga Cerrato, Jon Ølnes, Per Myrseth, and Inger-Mette Gustavsen. 2009. Time Challenges - Challenging Times for Future Information Search. *D-Lib Magazine*, 15(5/6).

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An International Standard for Semantic Annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Jannik Strötgen and Michael Gertz. 2013. Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*, 47/2.

Jannik Strötgen, Julian Zell, and Michael Gertz. 2013. HeidelTime: Tuning English and Developing Spanish Resources for TempEval-3. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 15–19, Atlanta, Georgia, USA. Association for Computational Linguistics.

Naushad Uzzaman, Hector Llorens, James F. Allen, Leon Derczynskiand, Marc Verhagen, and James Pustejovsky. 2012. TempEval-3: Evaluating Events, Time Expressions, and Temporal Relations. *CoRR*, abs/1206.5333.

Marc Verhagen, Robert Gaizauskas, Franck Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. SemEval-2007 - 15: TempEval Temporal Relation Identification. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, Prague, Czech Republic.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 Task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval'10)*, Uppsala, Sweden.