French Resources for Extraction and Normalization of Temporal Expressions with HeidelTime

Véronique MORICEAU and Xavier TANNIER
LIMSI-CNRS, Univ. Paris-Sud, Orsay, France,
firstname.lastname@limsi.fr

HeidelTime

- A multilingual, cross-domain temporal tagger according to the **TIMEX3** annotation standard
- A rule-based system with a separation between language-dependent resources and generic Java code:
  - **Resources:** extraction rules (regular expression patterns) and lexicons for normalization
  - **Extraction:** Absolute temporal expressions (*July 26th, 2013* ; *07-26-2013*) are extracted and normalized by the extraction rules
    Relative temporal expressions (*yesterday*, *July*) are extracted and left underspecified
  - **Normalization:** according to the type of documents (news, scientific, etc.) and to the tense of the verb used in the sentence

English
German
Dutch
Vietnamese
Arabic
Spanish
Italian
French

Jannik Strötgen and Michael Gertz. *Multilingual and Cross-domain Temporal Tagging*. In Language Resources and Evaluation, 269-298, 2013, Springer.

http://code.google.com/p/heideltime/

Development of French Resources

- **31 pattern files:** words and phrases used to express temporal expressions (months, days, etc.)
- **24 normalization files:** normalization information about the patterns (for example, the normalized value of *February* is *02*)
- **4 rule files:** date (100 rules), time (20 rules), duration (25 rules), set expressions (12 rules)

```
RULENAME="date_r1",
EXTRACTION="([Ll]e )%reWeekday %reDayNumber (%reMonthLong|%reMonthShort) %reYear4Digit",
NORM_VALUE="group(7)-%normMonth(group(4))-%normDay(group(3))"
```

jeudi 4 octobre 2012 → 2012-%normMonth(octobre)-%normDay(4) → 2012-10-04
le lundi 23 sept. 2013 → 2013-%normMonth(sept.)-%normDay(23) → 2013-09-23

```
RULENAME="date_r7",
EXTRACTION= "(%reMonthLong|%reMonthShort)", NORM_VALUE="UNDEF-year-%normMonth(group(1))"
```

DCT = 2009-09
Il est parti en mars (He left in March) → UNDEF-year-%normMonth(mars) + past tense → 2009-03-XX
Il reviendra en mars (He will come back in March) → UNDEF-year-%normMonth(mars) + future tense → 2010-03-XX

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.)*

Evaluation on the French TimeBank

https://gforge.inria.fr/projects/fr-timebank/

- **108 newspaper articles** in French annotated according to the ISO-TimeML standard
- **425 temporal expressions:**
  - 227 dates
  - 130 time expressions
  - 52 duration expressions
  - 16 temporal sets

	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		

	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 %

Evaluation on a User Application : automatic event timelines

Detection of salient dates in texts to automatically build event timelines from a search query:

- preprocessed newswire article corpus and normalization of temporal expressions
- indexation of the corpus by the Lucene search engine
- extraction of dates from retrieved documents
- ranking of dates to show the most important ones to the user

Corpus: 1 million newswire texts over the 2004-2011 period provided by the AFP French news agency
About 7% of absolute dates in the corpus

Performances with corpus preprocessed by HeidelTime or XIP (Xerox):
Comparison of runs with 94 manually-written chronologies according to Mean Average Precision

MAP with the corpus processed by XIP:	0.60
MAP with the corpus processed by HeidelTime:	0.64

R. Kessler, X. Tannier, C. Hagège, V. Moriceau, A. Bittar. *Finding Salient Dates for Building Thematic Timelines*. 50th annual meeting of the Association for Computational Linguistics (ACL 2012)

Error Analysis

French TimeBank:

- Adverbs *maintenant* (now), *aujourd'hui* (today) and *désormais* (henceforth) inconsistently annotated in the corpus:
  - type: either as TIME or DATE
  - value: either PRESENT-REF or normalized value

  → *22 mismatches* (DATE with HeidelTime)

- Dates associated with time expressions:
  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>

- Very specific expressions not in the pattern files:
  - time: time expressed in minutes or seconds
  - duration: *demi-siècle* (half-century), *quinquennat* (quinquennium), *période gréco-romaine* (greco-roman period)

  → *8 occurrences*

AFP corpus:

- Wrong tense identification:

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)

This work has been partially funded by French National Research Agency (ANR) under project Chronolines (ANR-10-CORD-010)