HeidelTime as a Baseline Temporal Tagger for All Languages

Jannik Strötegen and Michael Gertz
Database Systems Research Group, Heidelberg University, Im Neuenheimer Feld 348, 69120 Heidelberg, Germany

Main challenges [1]
- different domains, different challenges
- normalizing non-explicit expressions, e.g., today, Monday, next week, July
- only few languages addressed so far

Multilingual temporal tagging
- annotated corpora in several languages
- few multilingual temporal taggers
- earlier works on automatic extensions to new languages less successful

So far: temporal tagging of many languages never addressed!

Key features [1]
- rule-based system
- multilingual and domain-sensitive

Architecture
- language-independent, domain-specific normalization strategies
- language-dependent resources
- required: sentence, token, pos information

Language Resources
(i) pattern files
(ii) normalization files
(iii) rules

Example: "el 20 de enero de 2012" (2012-01-20)

Name="date_r1"
Extract="[Ee]l %reDayNum de %reMonthLong de %reYear4Digit"
Value="group(3)-%normMonth(group(2))-%normDay(group(1))"

Manual Extension to Languages

Procedure [2]
- linguistic preprocessing
- based on source language (English):
  1. manual translation of pattern and normalization files
  2. iterative rule development
  3. error analysis based improvements on annotated data (target language)

Disadvantages
- time- and labor-intensive
- language expert required

Automatic Extension of HeidelTime to All Languages

Strategy
- avoid language dependency
  → generic, simple sentence/token splitter
  → no pos tagger
- language-independent (A,C) and English translation-amendable (B) resources
  (A) → usable for all languages
  (B) → to automatically create pattern and normalization files for all languages
  (C) → rules without pos constraints and language-specific terms
    → 'creative' rules with wildcards
    → iterative improvements of (A,B,C)
      based on English annotated corpora

HeidelTime 2.0 at GitHub
- 13 languages (manual resources)
- 200+ languages (automatic resources)
- UIMA component
- Java standalone
- online demo

HeidelTime as temporal tagging baseline and starting point for 200+ languages!

References

This work was presented at EMNLP 2015, September 17–21, 2015, Lisbon, Portugal.