HeidelTime as A Baseline Temporal Tagger for All Languages

Jannik Strötgen 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!

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

Coverage and Evaluation

Evaluation
- annotated corpora of 10 languages
- comparison with manual HeidelTime
- manual resources work better
- results depend on Wiktionary coverage, language characteristics (morphology, token boundaries, ...)
- promising results for first baselines

Ongoing Work
- further translation resources
- more language-independent rules
- further fuzzy matching rules
- non-whitespace token boundaries

References

Availability

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!

Contact Information:
Jannik Strötgen
stroetgen@uni-hd.de
http://dbs.ifi.uni-heidelberg.de/

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