HeidelTime:
High Quality Rule-based Extraction and Normalization of Temporal Expressions

Jannik Strötgen and Michael Gertz

Institute of Computer Science, University of Heidelberg
Im Neuenheimer Feld 348, 69120 Heidelberg, Germany

Temporal Expressions (TEs)

Different types

- Date: On May 22, 1995, Farkas was ...
- Time: ... in Brownsville around 7:15 p.m.
- Duration: He spent six days abroad ...
- Set: ... for liver transplants each year ...

Different occurrences in documents

- explicit easy to normalize
- implicit knowledge is needed
- relative reference time is needed (& additional information)

Annotation scheme

- TimeML: ISO standard for temporal annotation (Timex3) [2]

Main Challenges

Disambiguation of:

- the reference time of relative TEs
- underspecified relative TEs
- relations between underspecified TEs and reference time (which Thursday?)

Document Creation Time: 1990-08-15
Also today, King Hussein of Jordan arrived in Washington ... he had rejected in peace talks following the August 1988 cease-fire ... which had condemned Iraq 's invasion of Kuwait on Aug. 2 ... The monarch will meet Bush on Thursday ... Independence Day 1989 ...

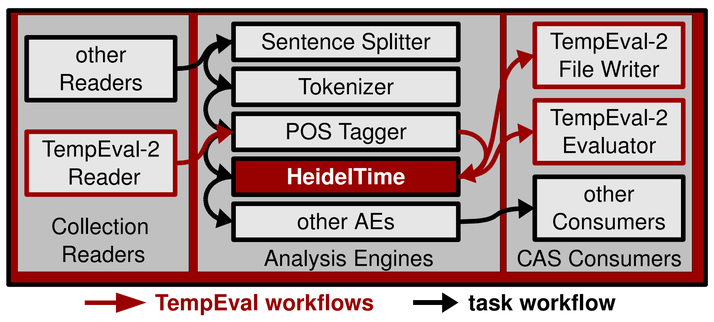
Task Description

- Identify extents of TEs
- Assign type information to TEs
- Normalize values according to Timex3

HeidelTime - Overview

- Rule-based system
- Realized as UIMA component [1]
- Methods for Extraction: regular expressions & NLP features
- Methods for Normalization: knowledge resources & linguistic clues
- Two rule sets: high precision (HeidelTime-1) and high recall (HeidelTime-2)

UIMA Text Mining Pipeline



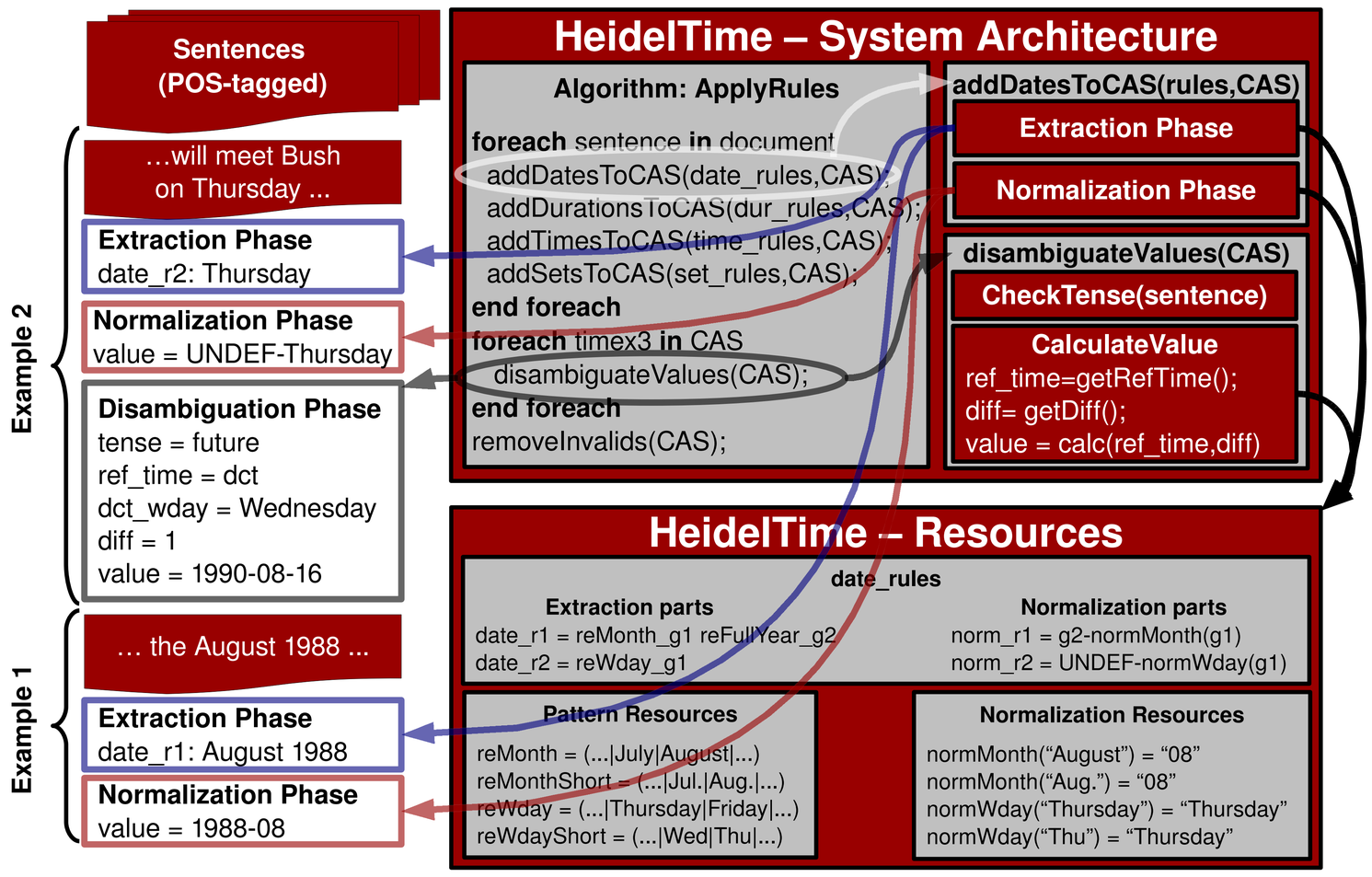
Iterative Rule Development Process

- Training data (sentence, token, Timex annotations → gold standard)
- Evaluator compares gold standard with HeidelTime Timex annotations
  Evaluator creates lists with FN, FP, TP
- These lists are used to improve rules

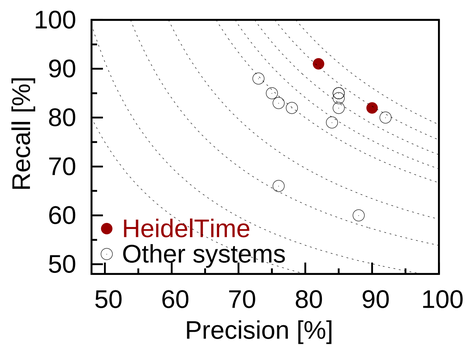
Evaluation Process

- Test data (sentence, token annotations)
- Files for evaluation are created

HeidelTime – System Architecture

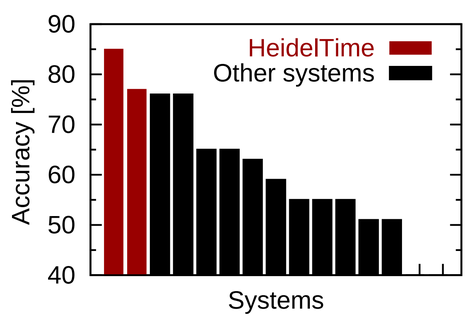


Extraction Results



	Precision	Recall	F-score
HeidelTime-1	90 %	82 %	86 %
HeidelTime-2	82 %	91 %	86 %

Normalization Results



	Value Acc.	Type Acc.
HeidelTime-1	85 %	96 %
HeidelTime-2	77 %	92 %

Ongoing Work

Other languages

- new pattern resources (e.g., names of months, weekdays, ...)
- new normalization resources
- new rules

Other types of corpora

- TempEval-2 corpus: news documents
- adaptations for and evaluation on other types of corpora

References

[1] UIMA, http://uima.apache.org/

[2] James Pustejovsky and Marc Verhagen: SemEval 2010 Task 13: Evaluating Events, Time Expressions, and Temporal Relations. In *SEW-2009*, pages 112-116, 2009.

Contact information:

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