Extending HeidelTime for Temporal Expressions
Referring to Historic Dates

Jannik Strötgen, Thomas Bögel, Julian Zell, Ayser Armiti, Tran Van Canh, Michael Gertz
Database Systems Research Group, Heidelberg University, Im Neuenheimer Feld 348, 69120 Heidelberg, Germany

Motivation – Temporal Tagging

Temporal Tagging
- extraction & normalization of temporal expressions

Main Challenge
- normalizing relative and underspecified expressions

Different Domains [1]
- pose different challenges
- require different strategies

Existing Approaches
- focus on English
- focus on news documents

Common Challenges
- temporal taggers referring to AD numbers:
  - need to handle
    - various “BC/AD” phrases
    - take care of tagging quality of news/current content
    - ambiguities:
      - day or year (April 20)
      - year or number (in 200 ...)
    - “BC/AD” phrases optional

Example of Tagging
- date: “in the year 90 ...”

→ context-dependent

The AncientTimes Corpus

Domain of Interest
- documents about history

Research Situation
- historic dates not well covered by temporal taggers (HeidelTime so far: only expressions after 1000 AD)
- historic dates hardly occur in existing corpora

<table>
<thead>
<tr>
<th>Date</th>
<th>Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1978</td>
<td>December 27</td>
</tr>
<tr>
<td>1998-04-18</td>
<td>today</td>
</tr>
<tr>
<td>2009-12-19</td>
<td>May 22, 1993</td>
</tr>
</tbody>
</table>

HeidelTime

Key Features [2]
- rule-based system
- multilingual
- domain-sensitive
- extraction: regular expressions & NLP features
- normalization: knowledge resources & linguistic clues

Language-independent
- resource interpreter

Language-dependent
- pattern files
  - month=.../April/May/...
- normalization files
  - normMonth(April)=04
- rule files

HeidelTime Adaptations

Pattern / Normalization Resources
- 1–4 digit years
- “BC/AD” phrases

Rules
- several modifications and new rules
- rules with explicit “BC/AD” phrases
- rules without “BC/AD” phrases

Normalization Strategies
- calculations for relative and underspecified expressions
- disambiguation of 2-digit years
- expressions without “BC/AD” phrases
  - preliminary normalized to AD dates
  - if explicit BC dates in documents, then check for chronology

Post-processing step is important

Evaluation

- same or better results on all publicly available corpora
- HeidelTime on AncientTimes (value F1)

<table>
<thead>
<tr>
<th>Language</th>
<th>Date</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>1.5 new</td>
<td>24.3</td>
<td>84.3</td>
<td>0.6</td>
</tr>
<tr>
<td>German</td>
<td></td>
<td>16.5</td>
<td>80.4</td>
<td>+1.8</td>
</tr>
<tr>
<td>Spanish</td>
<td></td>
<td>29.3</td>
<td>85.7</td>
<td>+0.5</td>
</tr>
<tr>
<td>Italian</td>
<td></td>
<td>25.4</td>
<td>81.1</td>
<td>+6.6</td>
</tr>
<tr>
<td>French</td>
<td></td>
<td>26.8</td>
<td>89.6</td>
<td>+24.6</td>
</tr>
<tr>
<td>Dutch</td>
<td></td>
<td>23.7</td>
<td>88.2</td>
<td>0</td>
</tr>
<tr>
<td>Arabic</td>
<td></td>
<td>25.3</td>
<td>81.9</td>
<td>+3.2</td>
</tr>
<tr>
<td>Vietnamese</td>
<td></td>
<td>15.6</td>
<td>89.4</td>
<td>0</td>
</tr>
</tbody>
</table>

* Increase due to post-processing

Availability

AncientTimes Corpus
- on HeidelTime page

HeidelTime’s Current Version
- UIMA component
- Java standalone
- online demo
- @ Google code

10 Languages
- English, German, Dutch, Spanish, Italian, French, Arabic, Vietnamese,
- new: Chinese, Russian

Ongoing Work
- further languages and improvements

References


This work was presented at LREC 2014, the 9th Language Resources and Evaluation Conference, May 26–31, 2014, Reykjavik, Iceland.

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