Chinese Temporal Tagging with HeidelTime

Hui Li, Jannik Strötgen, Julian Zell, and Michael Gertz
Database Systems Research Group, Heidelberg University,
Im Neuenheimer Feld 348, 69120 Heidelberg, Germany

Motivation

Temporal Taggers
- focus on English
- no Chinese temporal tagger is publicly available
- machine learning & rule-based approaches for temporal tagging

Our Contributions
- extend Heideltime to Chinese
- first publicly available Chinese temporal tagger
- improvements of Chinese TempEval-2 corpus

HeidelTime: a multilingual, cross-domain temporal tagger
- extraction: regular expressions & NLP features
- normalization: knowledge resources & linguistic clues

TempEval-2 Corpus

Annotation Standard: TimeML TIMEX3

Original Chinese Corpus
- many undefined value attributes
- several missing/wrong types
- some further errors

Re-annotation of the Corpus
- manually assign normalized values to undefined expressions
- correct existing expressions that have incorrect types

<table>
<thead>
<tr>
<th>corpus</th>
<th>temp. date / time / duration / set</th>
<th>undef. value</th>
</tr>
</thead>
<tbody>
<tr>
<td>training set</td>
<td>docs expr.</td>
<td>623 / 10 / 113 / 0 85</td>
</tr>
<tr>
<td>improved</td>
<td>765 628 / 125 / 2 0</td>
<td></td>
</tr>
<tr>
<td>test set</td>
<td>original</td>
<td>160 / 0 / 27 / 3 47</td>
</tr>
<tr>
<td>improved</td>
<td>193 166 / 23 / 4 0</td>
<td></td>
</tr>
</tbody>
</table>

Chinese HeidelTime Resources

Used Resources
- TempEval-2 training documents used for Chinese resource development

Four Steps to Add a New Language:
(1) Preprocessing:
- sentence, token, POS information
- HeidelTime uses TreeTagger [2]
- Chinese TreeTagger module available

(2) Translation of Pattern Files:

- \# reWeekday
<table>
<thead>
<tr>
<th>Monday</th>
<th>Tuesday</th>
<th>Wednesday</th>
</tr>
</thead>
<tbody>
<tr>
<td>星期一</td>
<td>礼拜一</td>
<td>周一</td>
</tr>
<tr>
<td>礼拜一</td>
<td>周一</td>
<td></td>
</tr>
<tr>
<td>周一</td>
<td></td>
<td></td>
</tr>
<tr>
<td>星期二</td>
<td>周二</td>
<td></td>
</tr>
<tr>
<td>礼拜二</td>
<td>周二</td>
<td></td>
</tr>
<tr>
<td>周二</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(3) Translation of Normalization Files:

- \# "normWeekday"
<table>
<thead>
<tr>
<th>Monday</th>
<th>Tuesday</th>
<th>Wednesday</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Monday&quot;</td>
<td>&quot;Tuesday&quot;</td>
<td>&quot;Wednesday&quot;</td>
</tr>
</tbody>
</table>

(4) Iterative Rule Development
- starting with (simple) English rules
- checking training corpus for errors
- adapting patterns, normalizations, and rules to improve results on training data

TempEval-2 Evaluation Results

Extraction & Normalization

- training set: P R F value type
  - original: 96.1 92.7 94.4 80 93
  - improved: 97.6 94.4 96.0 92 95

- test set: original 93.4 82.0 87.3 70 93
  - improved: 95.5 83.8 89.3 87 96

Normalization only

- training set: AU13-clean # correct
  - original: 90% 94% 97% 89% 95%
  - improved: 86% 96% 97% 96% 574

- test set: AU13 48% 87% 60% 97% 86
  - HeidelTime 70% 93% 89% 96% 121

Related Approach

Angeli and Uszkoreit [3]
- discriminative parsing approach
- language-independent
- normalization of temporal expressions
- no extraction of temporal expressions

Summary

Chinese Resources
- first publicly available Chinese temporal tagger of high quality
- available re-annotated Chinese TempEval-2 data sets

References

HeidelTime’s Current Version
- as UIMA component
- as standalone version (Java)
- online demo
- @ Google code

Languages
- English, German, Dutch, Spanish, Italian, Arabic, Vietnamese, Chinese, French

Availability

Contact Information:
Hui Li
hui.li@informatik.uni-heidelberg.de
http://dbs.ifi.uni-heidelberg.de/

This work was presented at EACL 2014, the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden.