HeidelTime: Temporal Tagging on Different Domains

Jannik Strötgen and Michael Gertz

Database Systems Research Group, Heidelberg University, Im Neuenheimer Feld 348, 69120 Heidelberg, Germany

Motivation

Temporal Tagging
• focus mainly on news documents
• only recently on narratives

Existing Annotated Corpora
• news: TimeBank
• narratives: WikiWars

Different Domains
• pose different challenges
• require different strategies

Newly Developed Corpora

Time4SMS
• 1,000 documents of NUS SMS corpus
• manually annotated
• total of 1,341 temporal expressions

Time4SCI
• 50 PubMed articles on clinical trails
• manually annotated
• total of 317 temporal expressions

Challenges on Different Corpora

News [1998-04-18]...
... for the United States, he said today,...
... On May 22, 1995, Farkas was made a brigadier general, and the following year,...
However, cited by police in December for driving under the influence of alcohol,...

• reference time often DCT
• relation to reference time
• many underspecified and relative expressions

Narrative [2009-12-19]
1979: Soviet invasion...
... land in Kabul on December 25,...
... they were complying with the 1978 Treaty of Friendship,...
... entered Afghanistan from the north on December 27. In the morning, the 103rd...

• reference time often in text
• relation to reference time
• long documents, rich discourse structure

SMS [2010-01-10] [5:19]
Whats ut u Wanted 2 say last nite?
SMS [2010-09-23] [9:50]
Yo! Rem 2 come for lab atm....
SMS [2011-02-16] [12:42]
... Andy is available at 10 am in his office

• relation to reference time
• non-standard language (errors, word creations,...)
• missing context information

Scientific [2009-12-19]
... Subjects consumed one tablet per day containing,...
... Subjects were assessed at baseline three and six months,...
Clinical pathology analysis was performed at baseline and six months...

• often no real reference time
• local semantics (document time frame)
• many durations and sets

HeidelTime’s Features [2]

• multilingual (English, German, Dutch)
• easy extensible to further languages
• cross-domain (news, narrative, colloquial, scientific)

Availability

HeidelTime
• as UIMA component
• as standalone version (Java)
• online demo
• @ Google code

Corpora
• Time4SMS
• Time4SCI
• WikiWarsDE

References


This work was presented at LREC 2012, the 8th Conference on Language Resources and Evaluation, May 21-27, 2012, Istanbul, Turkey.

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