Call for Master Thesis

FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

Temporal Signature Dataset from Wikipedia

The goal of this thesis is the creation of temporal signatures for all Wikipedia article texts. There are more than 6 mio articles in Wikipedia. Each article contains dates, as e.g. birth date and death date of persons, begin and end of events, etc. Dates can be extracted from the Wikipedia article texts via existing tools for temporal tagging, as e.g. HeidelTime [1].

For all articles a ∈ Wikipedia thereby a temporal signature $tsig_a$ is created as a vector

$tsig_a = (n_0, n_1, n_2, ..., n_{2018}, n_{2019}, n_{2020})$
with n_i = number of occurrences of year i in article a

The already existing method for the creation of temporal signatures now should be extended in the following way: besides extracting years directly from Wikipedia article a, also references of article a from other Wikipedia articles $\{b_1,...,b_n\}$ should be analyzed and years extracted from their context (as e.g. sentence or paragraph). I.e. for each article $b_i \in \{b_1,...,b_n\}$ that contains a link to article a, the already existing temporal signature $tsig_a$ should be extended with the number of occurrences of year j in article b_i but only from a limited context, i.e. the sentence containing the link to article a, or the paragraph containing the link to article a.

Task Description:

- Extract the Wikitext from the (english) Wikipedia Dumps [3] via WIkiExtractor [2]
- Extract original temporal signatures [1,4,5] (or use already computed signatures from [5])
- Extend original signatures with additional temporal information extracted from referencing (linking) documents (**NEW**) with HeidelTime temporal tagger [1]
  - Use sentence containing the link as context
  - Use paragraph containing the link as context
- Provide a dataset with extended temporal signatures for all Wikipedia articles.
- Provide the software in a reusable, well documented, and easy adjustable form
- AddOn: Try this also on other major language editions of Wikipedia (German, French, Dutch, etc.)

References:

[1] HeidelTime: https://github.com/HeidelTime/heideltime
[2] WikiExtractor: https://github.com/attardi/wikiextractor
[3] Wikimedia Dumps: https://dumps.wikimedia.org/
[4] Prabal Agarwal, Jannik Strötgen, Luciano del Corro, Johannes Hoffart, Gerhard Weikum: *diaNED: Time-Aware Named Entity Disambiguation for Diachronic Corpora* ACL 2018, https://drive.google.com/open?id=1G4Kta-A4r0sz80GDygbeb4xDjByQVtEA
[5] Resources for [4], timeNED.zip

Contact person:
Prof. Dr. Harald Sack
harald.sack@fiz-karlsruhe.de
harald.sack@kit.edu

aifb
Institute of Applied Informatics and Formal Description Methods
http://www.aifb.kit.edu/

Karlsruhe Institute of Technology