Building a Gold Standard for Temporal Entity Extraction from Medieval German Texts

Natalia Korchagina
1. Introduction
   1.1 Motivation
   1.2 Data
   1.3 Task

2. Creation of a Gold Standard Corpus
   2.1 Dataset
   2.2 Annotation settings
   2.3 Annotation guidelines
   2.4 Resulting Corpus

3. Experiments based on the Gold Standard

4. Future work
1.1. Introduction. Motivation.

- Importance of temporal dimension:
  - Undiscovered facts about historical events/personalities

- Manual extraction is time-consuming
- Needed: a tool for automatic temporal entity extraction

- Application examples of temporal tagging:
  - **information extraction:**
    - events are summarized and chronically ordered
  - **information retrieval:**
    - time as a query topic
1.1. Introduction. Motivation.

- Specificity of temporal expressions in historical texts: high spelling and lexical variation
- Use of tools for modern texts doesn’t lead to good results

<table>
<thead>
<tr>
<th>expression in old German</th>
<th>meaning in modern German</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xbris</td>
<td>Dezember</td>
</tr>
<tr>
<td>sant Conrats tag anno etc. lxxxxiij</td>
<td>26 November 1493</td>
</tr>
</tbody>
</table>

A temporal entity extraction system adapted for historical corpora is need.

**Why** do we need a Gold Standard?
- Base for quality estimation of the temporal extraction system
- No corpora of temporal annotations for historical German available
1.2. Introduction. Data.

- The project is funded and the data is provided by the Swiss Law Sources Foundation.

- Historical Swiss legal texts
- 118 volumes of critical editions available at [http://ssrq-sds-fds.ch/online/](http://ssrq-sds-fds.ch/online/)
- Available for digital processing:
  - 28 volumes
- ~7 million tokens of historical data
- 23,5 of 28 available volumes are in German
- Text creation time: 10th-18th centuries
- Multilinguality: texts in **German**, French, Latin, Romansh, Italian.
1.3. Introduction. Task.

- **Goal:** build a corpus of gold standard annotation of temporal entities in historical German texts.

**Workflow:**
1. Select data to be manually annotated.
2. Adapt annotation guidelines.
3. Analyse the annotated data.
4. Use the resulting corpus as a base for experiments.

- **Number of articles**: 50
- **Volumes**: 9 volumes from 5 cantons
- **Size**: about 32,000 tokens
- **Domain**: legal
- **Types of documents**: legal cases and transactional documents

**Period:** 1450-1550 (Early New High German): abundance of texts, old yet readable

Annotators: 2

Time allocated: 40 hours

Corpus given: set of articles automatically annotated with a rule-based temporal tagger HeidelTime (Strötgen und Gertz, 2013) adapted for Text+Berg corpus (Volk et al., 2010), (Rettich, 2013).

200 text segments were identified initially by HeidelTime as temp.expressions

Annotation framework: Oxygen XML Author with a CSS and an XML-schema adapted for the task.

391. Bern und Solothurn teilen die Herrschaften Bipp und Bechburg unter sich

[1463 April 29 (fritag vor dem meyen tag)]

Nachdem Solothurn vorgeschlagen hatte, die mit Bern gemeinsam verwalteten Herrschaften Bechburg und Bipp zu teilen, teilte Bern am 29. April 1463 mit, seine Gesandten, Venner Ludwig Hetzel und Ratsherr Urban von Muleren, hätten die Teile ausgeschieden; der Stadt Solothurn stehe die Wahl offen unter folgenden Bedingungen:

[1] Das ietweder herschaft fürwerthin zu ewigen ziten mit aller ir zügehörde an lüten, zinsen, nütten, nützen, verniessen, und güllen, als die harkomen ist, ein sundrige herschaft sin und bliben sol, mit hochen und niedern gerichten, twinger, veegen und stegen, als die in ietweder herschaft harkomen und gehalten worden sind.

2.3. Domain-specific annotation guidelines.

There are general guidelines for temporal and event annotations (Sauri et al., 2006)

Can’t be entirely reused:

- **too general**, don’t reflect particularities of historical corpora:
  
  developed for annotating instances of an English dataset of the TempEval 2010,
  
  don’t have old German-specific examples, e.g. with saint days.

- **too detailed**:

  describe the use of more attributes then we need for the task, “mod”, “temporalFunction”

What to annotate?

How to annotate?
2.3. Domain-specific annotation guidelines.

What to annotate?

Different kinds of temporal expressions to be considered:

**explicit:** 25. December 1490

**implicit:** Christmas

**relative:** Friday after Christmas

**markable:** today  
(can be situated on a timeline)

**non-markable:** soon  
(cannot be situated on a timeline)
2.3. Domain-specific annotation guidelines.

What to annotate?

TIMEX3 Extent:

1) General rule of span-economy: the extent of a tag should be as small as possible:

   \[ <\text{TIMEX3}>am\ 3.\ Oktober\ 1999</\text{TIMEX3}> \]

   \[ \text{vs.} \]

   \[ am\ <\text{TIMEX3}>3.\ Oktober\ 1999</\text{TIMEX3}> \]

2) Grammatical categories:

   - Noun, e.g. \[ \text{Montag} \]
   - Noun Phrase (NP), e.g. \[ \text{suntag\ judica\ in\ der\ vasten} \]
   - Adjective, e.g. \[ \text{jährliche\ <...>} \]
   - Adverb, e.g. \[ \text{nun} \]
   - Adjective or adverb phrases, e.g. \[ \text{alle\ nacht} \]
2.3. Domain-specific annotation guidelines.

How to annotate?

**Task:** check the automatically produced annotation, add new tags when necessary

**Annotation format:** subset of the *TimeML* l-ge (Pustejovsky et al, 2003):

- *TIMEX3* tags for temporal exp. and *SIGNAL* tags
- Attributes: *type* and *value*:
  - Types: DATE/ TIME/ SET/ DURATION
  - Value: ISO 8601 format for date and time

```xml
<TIMEX3 tid="t695" type="DATE" value="1493-07-05">fritag</TIMEX3>
<SIGNAL sid="s15">nach</SIGNAL>
<TIMEX3 tid="t1064" type="DATE" value="1493-07-04">Ulrici</TIMEX3>
<TIMEX3 tid="t1065" type="DURATION" value="P1D" beginPoint="t695" endPoint="t1064"/>
```

- Inter-annotator agreement:

<table>
<thead>
<tr>
<th></th>
<th>Detection</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average observed agreement</td>
<td>0.75</td>
<td>0.89</td>
</tr>
<tr>
<td>Cohen’s Kappa</td>
<td>0.74</td>
<td>0.76</td>
</tr>
</tbody>
</table>

- After adjudication performed with MAI adjudication tool (Stubbs, 2011):

  867 tags:
  - 807 temporal expressions
  - 60 signals

Comparison between the Gold Standard and its predecessors, i.e. annotations produced:

1) automatically by a rule-based temporal tagger HeidelTime (HT)

2) by human annotators A1 and A2

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>Correctly classified</th>
<th>Correctly normalised</th>
</tr>
</thead>
<tbody>
<tr>
<td>HT</td>
<td>0.26</td>
<td>0.96</td>
<td>96%</td>
<td>81%</td>
</tr>
<tr>
<td>A1</td>
<td>0.93</td>
<td>0.95</td>
<td>91%</td>
<td>84%</td>
</tr>
<tr>
<td>A2</td>
<td>0.88</td>
<td>0.90</td>
<td>95%</td>
<td>79%</td>
</tr>
</tbody>
</table>

A rule-based temporal tagger cannot handle (so far):

- spelling and lexical variation, e.g. expressions with the meaning “evening”:

  Abend  aubend  aebents  stübglogen
  abende  aubends  abentz  zenacht gessen
  abends  abentt  abendes  zenacht essen
  abendes  abentts  stübglogge  znacht essen
  abend  abent  stübgloggen  Nacht essen
  aubent  abents  Stübgloggen  nachtessen
  aubents  aebent  stübgloge  schlaff trunck

- complex temporal expressions with Saints’ days:

  e.g. fritag nechst nach sanntt Gallen tag -> 22 Oktober
  (en: Friday after St. Galls day)

Spelling normalisation to facilitate the use of existing temporal taggers:

1) Replace historical variants of temporal expressions with modern ones

2) Pass the **normalised corpus** to a temporal tagger

```plaintext
<...> viertzehnhundert und neun und fünfzig Jahre. <...>
```

```plaintext
<...> vierzehnhundert und neun und fünfzig Jahre. <...>
```

Temporal tagger

```plaintext
<...><TIMEX3>vierzehnhundert und neun und fünfzig Jahre</TIMEX3>. <...>
```

Spelling normalisation to facilitate the use of existing temporal taggers:

1) Replace historical variants of temporal expressions with modern ones

2) Pass the normalised corpus to a temporal tagger

3) Some improvement in temporal entity extraction should be observed, but perfect results are not expected:

   - single temporal expressions will be matched, but not if they are part of more complicated ones, for which the extraction rule is absent

     4. März 1647 can be recognised,

     1647 März 4 cannot be recognised

   - there is always a possibility that a test set will contain some expressions that do not have similar variants in resources

Successful cases of applying normalisation techniques for similar tasks:

Verb identification in Early Modern Swedish texts.

Method used: edit-based.

Improvement: 64.2% (unnorm) to 86.2% (Pettersson et al., 2013)

Method used: character-based statistical machine translation.

Improvement: 64.2% to 87.7% (Pettersson et al., 2014)

POS-tagging of Early New High German texts

Method used: edit-based

Improvement: by 2% (Logâcev et al., 2014)

To what extent can be improved the performance of a temporal tagger developed for modern texts by using normalisation as a pre-processing step?

Trial of spelling normalisation methods.

1. **Levenshtein-based** approach with context-sensitive weights (Pettersson et al., 2013)

   Experiment with the resources for German provided in the default version:

   *GerManC* corpus (1650-1780) (Scheible et al., 2011)

   **Normalisation accuracy**: 66\% \(^1\)

   *Evaluation of the temporal annotation produced by HeidelTime before (HT) and after (HT\_norm\_Lev) normalisation:*

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
<th>Correctly classified</th>
<th>Correctly normalised</th>
</tr>
</thead>
<tbody>
<tr>
<td>HT</td>
<td>0.26</td>
<td>0.96</td>
<td>96%</td>
<td>81%</td>
</tr>
<tr>
<td>HT_norm_Lev</td>
<td>0.27</td>
<td>0.89</td>
<td>96%</td>
<td>85%</td>
</tr>
</tbody>
</table>

\(^1\) Calculated for a manually normalised part (10\%) of the Gold Standard
4. Future work

- Further evaluation of various normalisation techniques;
- After the best-performing normalisation method is established:
  1) manually post-process a significant portion of the normalised output;
  2) apply a modern temporal tagger and evaluate the improvement of tagging a normalised corpus.

For expressions, not recognised as temporal even after normalisation:

- Try machine learning, e.g. character-based classifiers learning the “shape of the word” (Klein et al., 2003) and (Qi et al., 2014)
References


Thank you for your attention! 
Questions?