Corpus-Preparation with WebLicht for Machine-made Annotations of Examples in Philosophical Texts

Christian Lück

FernUniversität Hagen

October 1, 2019
Outline

1 Introduction

2 Architecture: WebLicht for Preprocessing

3 Machine-made Annotations of Examples
   - First Stage
   - Second Stage

4 Reproducible Results with WebLicht as a Service
Recent research in literary studies and philosophy has underlined the role of examples for the formation of knowledge (Ruchatz, Willer, and Pethes 2007; Schaub 2010; Lück et al. 2013; Güsken et al. 2018–).

Research on examples has remained in an exemplary mode.

I.e. single examples are commented in detail following hermeneutical methods.
Examples in Humanities Research

- Recent research in literary studies and philosophy has underlined the role of examples for the formation of knowledge (Ruchatz, Willer, and Pethes 2007; Schaub 2010; Lück et al. 2013; Gусken et al. 2018–).
- Research on examples has remained in an exemplary mode.
- I.e. single examples are commented in detail following hermeneutical methods.
- Reason: For research on large amounts of examples, there is no data set.
Examples in Humanities Research

- DFG funded research project *Das Beispiel im Wissen der Ästhetik (1750–1850)*, FernUniversität in Hagen
- focus on philosophy of aesthetics
- frequent use of examples (the tulip, the horse, the Alps, ...)
- frequent reflexions on the use of examples
- interesting aspects:
  - controversies on examples
  - effects of examples on fundamental conceptual distinctions (beauty of nature vs. beauty of art)
  - examples show that aesthetic judgments are governed by systems of knowledge, while authors say that in aesthetic judgments the (scientific) terms are suspended (example of the bat and the duckbill)
  - ...
Additional value of a larger dataset of examples

- present an inventory of examples
- make historical cuts (ger. *historische Längsschnitte*) that reveal the course of the frequency of examples over the researched period
  - emergence
  - boom
  - disappearance
- correlate trends in the philosophy of aesthetics with other discourses
  - travel literature of the 18th century
  - colonial discourse of the 19th centuries
From manual to machine-made annotations

- started with a database and a web form
  - results in collectanea without context
- proceeded with manual annotations in full texts
  - to time-consuming
  - to complex
- revised the model
- machine-made annotations based on a rule-based two-stage process
Using WebLicht for:
- sentence splitting
- tokenization
- lemmatization
- POS tagging
- (constituent parser)

R for text mining
- principles of tidy data
- one token per row

https://weblicht.sfs.uni-tuebingen.de
Machine-made Annotations of Examples

- non-trivial task
Machine-made Annotations of Examples

- non-trivial task
- diverse linguistic forms
• non-trivial task
• diverse linguistic forms
• may be marked with surface markers ("e.g."), but not mandatory
non-trivial task

diverse linguistic forms

may be marked with surface markers ("e. g."), but not mandatory

several examples may be stringed together
Machine-made Annotations of Examples

- non-trivial task
- diverse linguistic forms
- may be marked with surface markers ("e.g.") , but not mandatory
- several examples may be stringed together
- a single example may span a single word, a phrase, a sentence or even a paragraph
Examples in Aesthetics

Domain-specific observations:
- there is a single significant token
- we call it the **head of the example**
- it is a noun, a main verb or an adjective
- low to mid-range term frequency
Examples in Aesthetics

Domain-specific observations:

- there is a single significant token
- we call it the **head of the example**
- it is a noun, a main verb or an adjective
- low to mid-range term frequency

Can we exploit these observations?
A two-stage process

Stage 1  Find the *head of the example* in a sentence that has a surface marker ("e. g.")!
A two-stage process

Stage 1 Find the head of the example in a sentence that has a surface marker ("e. g.")!

Stage 2 Find non-marked examples based on the list of example heads returned by stage 1!
A two-stage process

Stage 1  Find the *head of the example* in a sentence that has a surface marker ("e. g.")!

- more than one candidate

Stage 2  Find non-marked examples based on the list of example heads returned by stage 1!
A two-stage process

Stage 1  Find the *head of the example* in a sentence that has a surface marker (“e. g.”)!
  - more than one candidate
  - problem of selection

Stage 2  Find non-marked examples based on the list of example heads returned by stage 1!
A two-stage process

Stage 1  Find the *head of the example* in a sentence that has a surface marker ("e. g."!)
  - more than one candidate
  - problem of selection

Stage 2  Find non-marked examples based on the list of example heads returned by stage 1!
  - A token, that has once been an example, does not have be an example throughout the corpus.
A two-stage process

Stage 1  Find the *head of the example* in a sentence that has a surface marker ("e.g.")!

- more than one candidate
- problem of selection

Stage 2  Find non-marked examples based on the list of example heads returned by stage 1!

- A token, that has once been an example, does not have be an example throughout the corpus.
- The probability of a token being an example decreases with increasing frequency of the same token in the text.
A two-stage process

Stage 1 Find the head of the example in a sentence that has a surface marker ("e. g.")!
- more than one candidate
- problem of selection

Stage 2 Find non-marked examples based on the list of example heads returned by stage 1!
- A token, that has once been an example, does not have be an example throughout the corpus.
- The probability of a token being an example decreases with increasing frequency of the same token in the text.
- problem of decision
Frist Stage

- Calculate the sum $h$ of weighted feature values for each token in a sentence with a surface marker. Let $t$ be the token, $S$ the sentence and $D$ the document, $f_i$ the features and $w_i$ the weights, then

$$h(t, S, D) = \sum_{i \in I} w_i f_i(t, S, D)$$ (1)
Frist Stage

- Calculate the sum $h$ of weighted feature values for each token in a sentence with a surface marker. Let $t$ be the token, $S$ the sentence and $D$ the document, $f_i$ the features and $w_i$ the weights, then

\[ h(t, S, D) = \sum_{i \in I} w_i f_i(t, S, D) \]  

(1)

- The token with maximum $h$ in the sentence is the *head of the example*. 

Frist Stage

- Calculate the sum $h$ of weighted feature values for each token in a sentence with a surface marker. Let $t$ be the token, $S$ the sentence and $D$ the document, $f_i$ the features and $w_i$ the weights, then

$$h(t, S, D) = \sum_{i \in I} w_i f_i(t, S, D) \quad (1)$$

- The token with maximum $h$ in the sentence is the head of the example.

- Evaluated features:
  - PoS tag
  - token frequency and lemma frequency
  - distance from surface marker (in tokens and commas)
  - direction (before or behind the surface marker)
First Stage – PoS tag

\[
f_{POS}(x) := \begin{cases} 
1 & \text{if } x \in \{\text{NE, FM}\} \\
0.8 & \text{if } x \in \{\text{NN}\} \\
0.5 & \text{if } x \in \{\text{VVINF, VVIZU, VVPP}\} \\
0.4 & \text{if } x \in \{\text{VVFIN}\} \\
0.2 & \text{if } x \in \{\text{VMINF}\} \\
0.1 & \text{if } x \in \{\text{VAINF}\} \\
0 & \text{otherwise}
\end{cases}
\]
First Stage – Token Frequency

An adaption of the augmented normalized term frequency (Salton and Buckley 1988) is used.

\[
    f_{tf}(t, D) := \begin{cases} 
    1 - c \frac{\#(t,D) - 1}{\max\{\#(t',D) \mid f_{POS}(t') > 0\} - 1} & \text{if } f_{POS}(t) > 0 \\
    0 & \text{otherwise} 
    \end{cases} 
\]  

(3)

where \(\#(t)\) is the number of times a token occurs in a document and \(c\) is a linearity factor with \(0 < c < 1\).
First Stage – Distance from Surface Marker

Let $l(S)$ be the maximum number of tokens before or after the surface marker in sentence $S$. Let $z(t, S)$ be the number of tokens in sentence $S$ between token $t$ and the surface marker.

$$f_{dt}(t, S) := 1 - \frac{z(t, S)}{l(S)} \quad (4)$$
First Stage – Direction

\[ f_{tf}(t, S) := \begin{cases} 
\frac{1}{4} & \text{if } t \text{ precedes the marker} \\
\frac{3}{4} & \text{otherwise} 
\end{cases} \] (5)
First Stage – Result

- 52 unambiguous example markers “z. B.” in Immanuel Kant’s *Critique of Judgment*
First Stage – Result

- 52 unambiguous example markers “z. B.” in Immanuel Kant’s *Critique of Judgment*
- manually assigned weights: $w_{POS} = 3$, $w_{tf} = 0$, $w_{lf} = 4$, $w_{dt} = 2$, and $w_{dc} = 6$: 

[Example list of words and phrases related to Kant's philosophy]
First Stage – Result

- 52 unambiguous example markers “z. B.” in Immanuel Kant’s Critique of Judgment
- manually assigned weights: $w_{POS} = 3$, $w_{tf} = 0$, $w_{lf} = 4$, $w_{dt} = 2$, and $w_{dc} = 6$:
First Stage – Result

- 52 unambiguous example markers “z. B.” in Immanuel Kant’s *Critique of Judgment*
- manually assigned weights: \( w_{POS} = 3, \ w_{tf} = 0, \ w_{lf} = 4, \ w_{dt} = 2, \) and \( w_{dc} = 6: \)
- 5 errors out of 52
First Stage – Perspectives

- keep manually corrected list of example heads
- use this list for assigning the weights by regression
- lesson learned: Manual annotations for training ML algorithms for this very well defined task would have been simple.
Second Stage

- Results from stage 1 underline that stage 2 is non-trivial:
  - “Körper” occurs 33 times throughout the text, but only sometimes it is an example.
  - “Größe” occurs 48 times.
  - many of such abstract concepts (we call them pseudo examples)
Second Stage

- Results from stage 1 underline that stage 2 is non-trivial:
  - “Körper” occurs 33 times throughout the text, but only sometimes it is an example.
  - “Größe” occurs 48 times.
  - many of such abstract concepts (we call them pseudo examples)
- Task: define criteria (features) for decision
  - similarity of semantic contexts (occurrence of same tokens)
  - similarity of syntactic contexts
  - a frequency threshold for examples
  - ...

Task: define decision rules

Instead: manual annotations for training ML algorithms (e.g. decision tree learning)

Well defined and simple task for manual annotations:
- Search all occurrences of a given token (or lemma) and annotate, whether it is an example or not!
Second Stage

- Results from stage 1 underline that stage 2 is non-trivial:
  - “Körper” occurs 33 times throughout the text, but only sometimes it is an example.
  - “Größe” occurs 48 times.
  - many of such abstract concepts (we call them pseudo examples)

- Task: define criteria (features) for decision
  - similarity of semantic contexts (occurrence of same tokens)
  - similarity of syntactic contexts
  - a frequency threshold for examples
  - ...

- Task: define decision rules
Second Stage

- Results from stage 1 underline that stage 2 is non-trivial:
  - “Körper” occurs 33 times throughout the text, but only sometimes it is an example.
  - “Größe” occurs 48 times.
  - many of such abstract concepts (we call them pseudo examples)

- Task: define criteria (features) for decision
  - similarity of semantic contexts (occurrence of same tokens)
  - similarity of syntactic contexts
  - a frequency threshold for examples
  - ...

- Task: define decision rules

- Instead: manual annotations for training ML algorithms (e.g. decision tree learning)
Second Stage

- Results from stage 1 underline that stage 2 is non-trivial:
  - “Körper” occurs 33 times throughout the text, but only sometimes it is an example.
  - “Größe” occurs 48 times.
  - many of such abstract concepts (we call them pseudo examples)
- Task: define criteria (features) for decision
  - similarity of semantic contexts (occurrence of same tokens)
  - similarity of syntactic contexts
  - a frequency threshold for examples
  - ...
- Task: define decision rules
- Instead: manual annotations for training ML algorithms (e.g. decision tree learning)
  - Well defined and simple task for manual annotations:
Second Stage

- Results from stage 1 underline that stage 2 is non-trivial:
  - “Körper” occurs 33 times throughout the text, but only sometimes it is an example.
  - “Größe” occurs 48 times.
  - many of such abstract concepts (we call them pseudo examples)

- Task: define criteria (features) for decision
  - similarity of semantic contexts (occurrence of same tokens)
  - similarity of syntactic contexts
  - a frequency threshold for examples
  - ...

- Task: define decision rules

- Instead: manual annotations for training ML algorithms (e.g. decision tree learning)
  - Well defined and simple task for manual annotations:
  - Search all occurrences of a given token (or lemma) and annotate, whether it is an example or not!

Christian Lück (FernUniversität Hagen)  Corpus-Preparation with WebLicht for Machine Learning  October 1, 2019  18 / 22
Reproducible Results with WebLicht as a Service

```
curl -X POST -F chains=@$chain -F content=@$1 -F apikey=$WEBLICHTKEY $url
```
Reproducible Results with WebLicht as a Service

- Do not point and click! Use WebLicht as a Service and script your preprocessing tasks!
Reproducible Results with WebLicht as a Service

- Do not point and click! Use WebLicht as a Service and script your preprocessing tasks!
- Scripting makes preprocessing reproducible.
Reproducible Results with WebLicht as a Service

<table>
<thead>
<tr>
<th>WebLicht as a Service</th>
<th>NLP-libraries in Java, Python etc.</th>
</tr>
</thead>
</table>

Christian Lück (FernUniversität Hagen)  Corpus-Preparation with WebLicht for Machine Learning  October 1, 2019  21 / 22
### Reproducible Results with WebLicht as a Service

<table>
<thead>
<tr>
<th>WebLicht as a Service</th>
<th>NLP-libraries in Java, Python etc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>scripting</td>
<td>scripting</td>
</tr>
</tbody>
</table>
### Reproducible Results with WebLicht as a Service

<table>
<thead>
<tr>
<th>WebLicht as a Service</th>
<th>NLP-libraries in Java, Python etc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>scripting</td>
<td>scripting</td>
</tr>
<tr>
<td>preprocessing remains independent of the implementation of your model</td>
<td>tie preprocessing to the implementation of your model</td>
</tr>
<tr>
<td>WebLicht as a Service</td>
<td>NLP-libraries in Java, Python etc.</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-----------------------------------</td>
</tr>
<tr>
<td>scripting</td>
<td>scripting</td>
</tr>
<tr>
<td>preprocessing remains independent of the implementation of your model</td>
<td>tie preprocessing to the implementation of your model</td>
</tr>
<tr>
<td>concentrate on the specification of your model (not its implementation)</td>
<td>?</td>
</tr>
<tr>
<td>WebLicht as a Service</td>
<td>NLP-libraries in Java, Python etc.</td>
</tr>
<tr>
<td>-----------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>scripting</td>
<td>scripting</td>
</tr>
<tr>
<td>preprocessing remains independent of the implementation of your model</td>
<td>tie preprocessing to the implementation of your model</td>
</tr>
<tr>
<td>concentrate on the specification of your model (not its implementation)</td>
<td>?</td>
</tr>
<tr>
<td>no versioning</td>
<td>can be pinned to a specific version</td>
</tr>
</tbody>
</table>
## Reproducible Results with WebLicht as a Service

<table>
<thead>
<tr>
<th>WebLicht as a Service</th>
<th>NLP-libraries in Java, Python etc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>scripting</td>
<td>scripting</td>
</tr>
<tr>
<td>preprocessing remains independent of the implementation of your model</td>
<td>tie preprocessing to the implementation of your model</td>
</tr>
<tr>
<td>concentrate on the specification of your model (not its implementation)</td>
<td>?</td>
</tr>
<tr>
<td>no versioning</td>
<td>can be pinned to a specific version</td>
</tr>
<tr>
<td>changelog?</td>
<td>changelog present for most libraries</td>
</tr>
</tbody>
</table>


Güsen, Jessica et al., eds. 2018–. *z. B. Zeitschrift zum Beispiel*.


