»A Buster Keaton of Linguistics«: First Automated Approaches for the Extraction of Vossian Antonomasia

Michael Schwab, Robert Jäschke, Frank Fischer, Jannik Strötgen

**Background**

Vossian Antonomasia (VA)

- Rheto-ric/stylistic device
- Special case of Antonomasia, similar to Metonymy
- First discovered by Dutch humanist Gerardus Vossius
- Attributing a particular property to a person by naming another person

Source → Modifier → Target

Example: Before then, Einstein was already esteemed by many physicists as the Newton of the 20th century.

This work is licensed under a Creative Commons Attribution 4.0 International License.

**Corpus**

**Corpus creation**

Extends the approach of Fischer and Jäschke (2019)

1. Limit candidates by using regular expressions:
   \[ \text{regex WD-L} \]
   \[
   \b(\text{the}|an?)\s+(\{\w.,'-\}\s+){1,10}(\text{of}|\text{for}|\text{among})\b
   \]
   
2. Extract all names and aliases of entities of type »human« from Wikidata

3. Use popularity measure (e.g. #sitelinks) to remove candidates whose non-human counterpart is more popular

4. Extract a list with names and aliases containing the words »for, from, of«

5. Remove candidate, if there is a match

**Candidate Generation**

1. Limit candidates by using regular expressions:
   \[ \text{regex WD-L} \]
   \[
   \b(\text{the}|an?)\s+(\{\w.,'-\}\s+){1,10}(\text{of}|\text{for}|\text{among})\b
   \]

2. Extract all names and aliases of entities of type »human« from Wikidata

3. Match candidates with list

**Entity Recognition / Linking**

1. Use Stanford NER tagger to identify entities of type »person«

2. Extract all names and aliases of all entities from Wikidata

3. Use popularity measure (e.g. #sitelinks) to remove candidates

4. Extract a list with names and aliases containing the words »for, from, of«

5. Remove candidate, if there is a match

**Classification**

1. Train a bidirectional LSTM neural network with the corpus data to get a binary classifier

**Approaches**

<table>
<thead>
<tr>
<th>Approach</th>
<th>prec</th>
<th>rec</th>
<th>f1</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td></td>
<td></td>
<td>49.8%</td>
</tr>
<tr>
<td>WD</td>
<td>67.3%</td>
<td>93.0%</td>
<td>78.1%</td>
</tr>
<tr>
<td>NER</td>
<td>71.8%</td>
<td>81.3%</td>
<td>76.2%</td>
</tr>
<tr>
<td>BLSTM</td>
<td>86.9%</td>
<td>85.3%</td>
<td>86.1%</td>
</tr>
</tbody>
</table>

**Methods**

**Goal:** automated approach to identify VA in large text corpora

**Base Methods**

**Candidate Generation**

1. Limit candidates by using regular expressions:
   \[ \text{regex WD-L} \]
   \[
   \b(\text{the}|an?)\s+(\{\w.,'-\}\s+){1,10}(\text{of}|\text{for}|\text{among})\b
   \]

2. Extract all names and aliases of entities of type »human« from Wikidata

3. Use popularity measure (e.g. #sitelinks) to remove candidates whose non-human counterpart is more popular

**Entity Recognition / Linking**

1. Use Stanford NER tagger to identify entities of type »person«

2. Extract all names and aliases of all entities from Wikidata

3. Use popularity measure (e.g. #sitelinks) to remove candidates

4. Extract a list with names and aliases containing the words »for, from, of«

5. Remove candidate, if there is a match

**Classification**

1. Train a bidirectional LSTM neural network with the corpus data to get a binary classifier

**Approaches**

<table>
<thead>
<tr>
<th>Approach</th>
<th>prec</th>
<th>rec</th>
<th>f1</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td></td>
<td></td>
<td>49.8%</td>
</tr>
<tr>
<td>WD</td>
<td>67.3%</td>
<td>93.0%</td>
<td>78.1%</td>
</tr>
<tr>
<td>NER</td>
<td>71.8%</td>
<td>81.3%</td>
<td>76.2%</td>
</tr>
<tr>
<td>BLSTM</td>
<td>86.9%</td>
<td>85.3%</td>
<td>86.1%</td>
</tr>
</tbody>
</table>

**Statistics**

**Top 10 Sources**

1. Michael Jordan
2. Rodney Dangerfield
3. Johnny Appleseed
4. Babe Ruth
5. Elvis Presley
6. Michelangelo
7. Donald Trump
8. Madonna
9. Bill Gates
10. Pablo Picasso

**Top 10 Modifiers**

1. his day
2. his time
3. the 90’s
4. our time
5. baseball
6. his generation
7. tennis
8. her time
9. - the Newton of the 20th century
10. - the Prince of Wales

**Contact:**

Michael Schwab
email: michel.schwab@hu-berlin.de

Data and Code: https://vossanto.weltliteratur.net/

See https://vossanto.weltliteratur.net/theof/humans/statistics.html#top-40-gallery for detailed image credits