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

Michel Schwab, Robert Jäschke, Frank Fischer, Jannik Strötgen

Background

Vossian Antonomasia (VA)

Gerhard Johannes Vossius (1577–1649)

- Rhetoric/stylistic device
- Special case of **Antonomasia, similar to Metonymy**
- First discovered by Dutch humanist Geradus 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**.

Corpus

»The New York Times«, 1987 - 2007

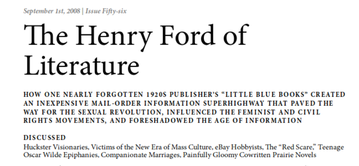
1 854 726 articles

Corpus creation

Extends the approach of Fischer and Jäschke (2019)

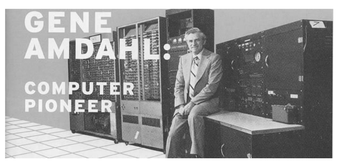
1. **Regex**: \b(the|an?)\s+([\w.,'-]+\s+){1,10}?(of|for|among)\b/
2. **Wikidata** entity matching (type »human«)
3. Handcrafted **blacklist**
4. Manual **labeling** by expert

pattern	regex	Wikidata	blacklist	true VA
the-of	12,748,735	90,712	3,591	2,779
a-of	5,900,839	11,860	705	118
an-of	956,247	4,539	88	14
the-for	2,960,459	8,070	817	24
a-for	1,869,946	4,812	536	59
an-for	304,529	1,424	296	13
the-among	122,345	139	13	3
a-among	67,019	82	25	13
an-among	11,158	12	1	0
sum	24,941,277	121,650	6,072	3,023



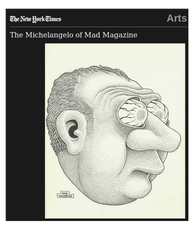
= Emanuel Haldeman-Julius (believermag.com, 2008)

Gene Amdahl: The Thomas Edison of Computing



= Gene Amdahl (blog.syncsort.com, 2016)

= Jim Koch, (theatlantic.com, 2014)



= Basil Wolverton (NYT, 2009)

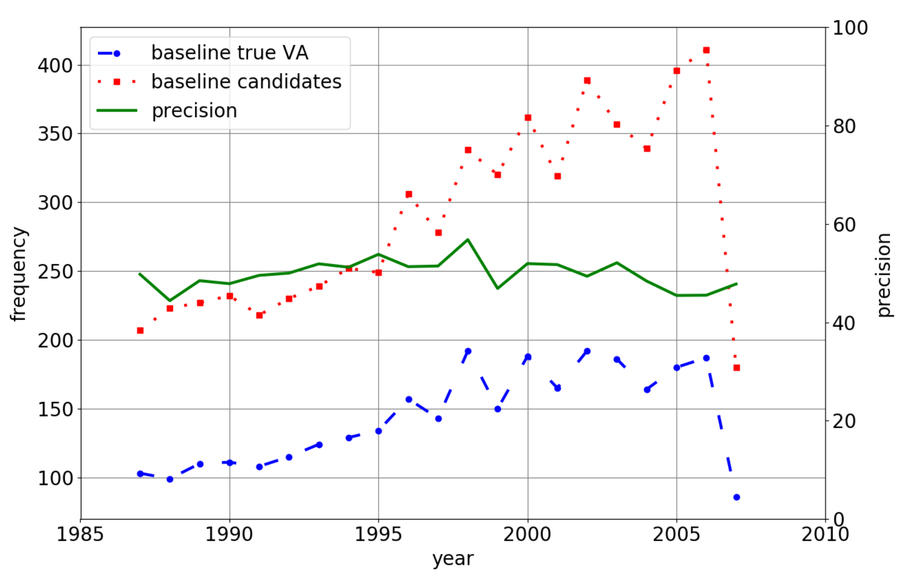
Statistics

Top 10 Sources

#	
71	Michael Jordan
61	Rodney Dangerfield
40	Johnny Appleseed
37	Babe Ruth
36	Elvis Presley
25	Michelangelo
25	Donald Trump
23	Madonna
23	Bill Gates
23	Pablo Picasso

Top 10 Modifiers

#	
56	his day
34	his time
29	Japan
20	the 90's
17	our time
17	China
16	baseball
16	his generation
16	tennis
14	her time



Methods

Goal: automated approach to identify VA in large text corpora

Base Methods

Candidate Generation

Regex

1. Limit candidates by using regular expressions:

```
\b(the|an?)\s+([\w.,'-]+\s+){1,10}?(of|for|among)\b
```

- the Robert Downey, Jr. of ✅
- a Shaquille O'Neal for ✅
- an Alfred Hitchcock among ✅
- the chinese Michael Jordan ❌

Entity Recognition / Linking

WD-L

1. Extract all names and aliases of entities of type »human« from Wikidata
2. Match candidates with list

WIKIDATA »instance of« (P31) »human« (Q5)

NER

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

WD-P

1. Extract all names and aliases of all entities from Wikidata
2. Use popularity measure (e.g. #sitelinks) to remove candidates whose non-human counterpart is more popular

»the House of« ?

WD-F

1. Extract a list with names and aliases containing the words »for, from, of«
2. Match source + multiple words that appear after the source with the list
3. Remove candidate, if there is a match

»the Prince of Wales « ?

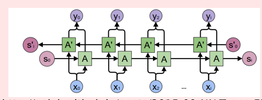
Expert

1. Apply a handcrafted blacklist

Classification

BLSTM

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



http://colah.github.io/posts/2015-09-NN-Types-FP/

Approaches

baseline	Regex	→	WD-L	→	Expert	
WD	Regex	→	WD-L	→	WD-P	→ WD-F
NER	Regex	→	NER	→	WD-F	
BLSTM	Regex	→	WD-L	→	Expert	→ BLSTM

Results

approach	prec	rec	f1
baseline	49.8%	—	—
WD	67.3%	93.0%	78.1%
NER	71.8%	81.3%	76.2%
BLSTM	86.9%	85.3%	86.1%

The Berlin Times

https://vossanto.weltliteratur.net/ Thursday, Nov 7, 2019 - Since 2019 -

BREAKING NEWS!

So far, BLSTM is the Michael Jordan of VA extraction

By M.Schwab, R. Jäschke, F. Fischer and J. Stroetgen

Attributing a particular property to a person by naming another person, who is typically wellknown for the respective property, is called a Vossian Antonomasia (VA).

This subtype of metonymy, which overlaps with metaphor, has a specific syntax and is especially frequent in journalistic texts. While identifying Vossian Antonomasia is of particular interest in the study of stylistics, it is also a source of errors in relation and fact extraction as an explicitly mentioned entity occurs only metaphorically and should not be associated with respective contexts. Despite rather simple syntactic variations, the automatic extraction of VA was never addressed as yet since it requires a deeper semantic understanding of mentioned entities and underlying relations.

In this paper, we propose a first method for the extraction of VAs that works completely automatically. Our approaches use named entity recognition, distant supervision based on Wikidata, and a bi-directional LSTM for postprocessing.The evaluation on 1.8 million articles of the New York Times corpus shows that our approach significantly outperforms the only existing semi-automatic approach for VA identification by more than 30 percentage points in precision.

Contact: Michel Schwab
email: michel.schwab@hu-berlin.de

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