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

Michel Schwab,[1] **Robert Jäschke,**[1,2] **Frank Fischer,**[3] **and Jannik Strötgen**[4]

[1]Humboldt University, Berlin, Germany

[2]L3S Research Center, Hannover, Germany

[3]National Research University Higher School of Economics, Moscow, Russia

[4]Bosch Center for Artificial Intelligence, Renningen, Germany

`{michel.schwab,robert.jaeschke}@hu-berlin.de`
`ffischer@hse.ru`
`jannik.stroetgen@de.bosch.com`

Abstract

Attributing a particular property to a person by naming another person, who is typically well-known for the respective property, is called a *Vossian Antonomasia* (VA). This subtpye 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.

1 Introduction

Background. Vossian Antonomasia (VA) is a stylistic device which attributes a certain property to a person by naming another (more well-known, more popular) person as a reference point. For instance, when Jim Koch is described as "the Steve Jobs of Beer" (Fallows, 2014), certain qualities of Steve Jobs, be it entrepreneurship or persuasiveness, are assigned to Jim Koch, co-founder and chairman of the Boston Beer Company.

VA is named after Gerardus Vossius (1577–1649), the Dutch classical scholar and author of rhetorical textbooks. Although the phenomenon is traceable back to antiquity (as shows the example of Crassus, who used to be ironically referred to as "the Palatine Venus"), it was first distinguished and described as a separate phenomenon by Vossius.

It constitutes a sub-phenomenon of the classic antonomasia. In the classification of Holmqvist and Płuciennik (2010) it is called *metaphorical antonomasia* or *antonomasia₂* and is described as a comparison "with paragons from other spheres". VAs are particularly popular in journalistic texts.

Definition. VAs consist of three parts: a *source* (in our example "Steve Jobs") serves as paragon to elevate the *target* ("Jim Koch") by applying a *modifier* ("of beer") that provides the corresponding context (see Bergien, 2013). Although in most cases both the source and the target are persons, one or both of them could be almost anything as long as they bear a proper name, for instance, software, as shows this example: "Word [...] was dubbed the Marquis de Sade of word processors, which was not altogether unfair." (NYT 1993/09/26/0636952)[1] The target is not necessarily part of the sentence (in particular when a VA is used in a headline – like in the title of this paper[2]) and it can be hypothetical, as in "Who is the Tolstoy of the Zulus?" (NYT 1988/01/03/0106769)

Challenges. Despite some interest in VAs in the humanities (e.g., Holmqvist and Płuciennik, 2010; Bergien, 2013), works on their automatic extraction are scarce. One reason could be that "antonomasia is essentially humanistic and anti-computational" (Holmqvist and Płuciennik, 2010) and thus identifying and understanding it in its entirety is often difficult even for humans and requires deep cultural

---

[1]To avoid an excessively long reference list, all examples taken from the New York Times corpus (Sandhaus, 2008) are cited using the pattern "NYT year/month/day/article-id".

[2]To resolve this example: "a Buster Keaton of linguistics" refers to the performance artist Stuart Sherman (NYT 1987/07/19/0057747).

background knowledge. For example, understanding the phrase "It's the Dolly Parton of cakes" (NYT 2007/02/14/1826062) requires specific knowledge about Dolly Parton, her skills or peculiarities. Only in rare cases like this, an explanation of the intended meaning ("a little bit tacky, but you love her") is provided. It is usually left to the reader to make sense of a VA.

Importance. From a humanities perspective, a VA constitutes an interesting phenomenon of enculturation (Holmqvist and Płuciennik, 2010) that deserves to be studied more in-depth, based on larger corpora.

However, we expect that the large-scale identification of VAs and other such stylistic devices is not only important in the humanities and for cultural reasons, but also in natural language processing to avoid mistakes in tasks such as machine translation and fact extraction. For instance, given the sentence "Today, the German Ronaldo quit his career.", we do not want to extract a fact that Ronaldo quit his career, but we need to detect that "the German Ronaldo" is a VA referring to a German soccer player.

In addition, entity disambiguation and coreference resolution could be improved. Consider the example "Jimmy Johnson is the Madonna of college football ..." (NYT 1987/01/02/0000431). We would like to refer "the Madonna of college football" to Jimmy Johnson, and, in particular, we want to avoid that Madonna is part of any coreference chain.

It could also help in new interesting question answering tasks, for example, "Who is the Bill Gates of Japan?", where the sentence "... Mr. Horie has made headline news with a success story that has turned him into the Bill Gates of Japan." (NYT 2006/01/19/1733197) could be one answer to this question.

Generally speaking, successful detection and resolution of VAs could be a further step towards full natural language understanding.

Contribution. We propose the first effective fully automated approaches for extracting VAs from texts. In our evaluation, we compare them against a baseline and assess the difficulty of identifying VAs using crowdsourcing. In addition, we extend the only existing larger data set (Fischer and Jäschke, 2019) with VAs of eight additional patterns. We have also double-checked and corrected existing

and new annotations. Our code and data are freely[3] available[4] and easily extensible towards other patterns and entity types.

2 Related Work

Very similar to metaphor detection (Tsvetkov et al., 2014; Gao et al., 2018; Mao et al., 2018), identifying VAs is a difficult task. So far, we are dealing with manually curated collections like the Wall Street Journal's collection of occurrences of "Michael Jordans of ..." (Cohen et al., 2015). The only attempt automating the extraction has been presented by Fischer and Jäschke (2019). Their semi-automatic approach for English texts extracts VA candidates for only one single pattern, *the ENTITY of*. Using a regular expression to identify candidates, names and aliases of Wikidata entities whose 'instance-of' property is 'human' as a first filter, and a manually created blacklist as second filter, they extract about 3,700 VA candidates. 70% of them have been confirmed as VAs by a domain expert.

Since the phenomenon is a kind of metonymy, its detection is closely related to approaches to metonymy resolution. This has been covered extensively, specifically by Markert and Nissim, for example, in (Nissim and Markert, 2003). They compare this task to word-sense disambiguation and propose a thesaurus-based classification approach. Figures of speech that have been covered include, for instance, metaphors (Tsvetkov et al., 2014; Gao et al., 2018; Mao et al., 2018) and irony (Akhtyrska, 2014). The supervised approach of Tsvetkov et al. (2014), for instance, can quite reliably "discriminate whether a syntactic construction is meant literally or metaphorically" but is very specific, focused on subject–verb–object and adjective–noun phrases.

3 Corpus Creation and Baseline

To create a corpus of candidate phrases and manually confirmed VAs, we use the dataset of Fischer and Jäschke (2019) as starting point and their method as baseline. However, instead of relying only on one individual syntactic pattern, *the ENTITY of*, we extended the candidate generation process with the patterns *the ENTITY among*, *the ENTITY for*, and six variations thereof where *the* is

---

[3]GPL3 and CC-BY, respectively

[4]https://vossanto.weltliteratur.net/ – in addition to our code and data, information about some of the authors' prior work on VAs is also available.

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

Table 1: Number of VA candidates after each step and manually confirmed VAs for all used patterns.

replaced by *a* or *an*, respectively. We identified these patterns to be the most prevalent patterns in a preliminary analysis of a manually maintained collection of VAs.

Using the New York Times corpus (Sandhaus, 2008) of 1,854,726 articles published between 1987 and 2007, the regular expression based pattern matching resulted in almost 25 million candidate phrases vs. less than 13 million candidate phrases in Fischer and Jäschke (2019)'s work.

Using Wikidata for distant supervision, we keep only candidates whose string matches the exact name or alias of Wikidata entities, respecting the case of letters, with 'instance-of' property 'human', and exclude candidates using the manually curated blacklist[5] provided by Fischer and Jäschke (2019).[6] We consider this procedure of "regex – Wikidata – blacklist" as baseline. In Table 1, we show the number of VA candidate phrases after each step.

In addition, all remaining candidates were manually annotated by a domain expert with 10 years of experience in the research of VAs and samples were double-checked by a trained student. The resulting corpus contains 6,072 VA candidates, 3,023 of them were classified to be true VAs by the human annotators. That is, the baseline approach has a precision of 49.8% (see Table 1). Note that due to the sparseness of the phenomenon, we did not try to estimate the recall of the baseline approach.

4 VA Extraction Approaches

We aim at a fully automated extraction of VAs from texts. For this, we develop two approaches to replace the need of a manually curated blacklist of the baseline approach using named entity recognition as well as Wikidata and a link-based popularity measure. In addition, as a further approach, we extend the baseline by classifying VA candidates using a bi-directional LSTM.

4.1 Wikidata (WD)

This is a modification of the baseline approach to avoid the manual curation of a blacklist. Thus, after the regex step and linking the source candidates with the Wikidata list of humans as in the baseline approach, we assess their popularity within Wikidata.

The rationale is that a prerequisite for being a source of a VA expression is to be famous or popular or notorious for something. Otherwise, the VA is less likely to be understood by a reader. This arguably goes along with being present on several of the almost 300 international Wikipedia versions. Therefore, we remove source candidates whose name also matches the label or alias of a non-human Wikidata entity that has more sitelinks[7] (i.e., is more popular according to our measure) than the linked human. For instance, the candidate 'the House of', that matched the American botanist Homer Doliver House[8] in the first step, will be removed since in the second step the entity 'house'[9] (the building, not the botanist) has more sitelinks (178 > 9). The popularity measure could easily be changed to any other measure like statements or using Wikipedia clickstream data.

In addition, we remove all candidates where the source (e.g., 'Prince') together with multiple words following it (e.g., 'of Wales') match the name or alias of another Wikidata entity (https://www.wikidata.org/wiki/Q43274). This allows us to remove frequent false positives like 'Prince of Wales'.

4.2 Named Entity Recognition (NER)

This approach is built on the regex step from the baseline approach as well, but we replace the way of identifying 'persons': instead of restricting the extraction to VAs with Wikidata entities, we per-

form Named Entity Recognition on the sentence candidates using the Stanford three-class named entity tagger (Finkel et al., 2005).

When all words of the candidate source (i.e., the words between the first and the last word of the regular expression match) are tagged as 'PERSON', we consider the candidate a positive match. However, to avoid false positives, we apply the last step from the WD approach (Sec. 4.1) and remove again all candidates whose source together with following words match another Wikidata entity.

4.3 Bi-directional Long Short-Term Memory Neural Network (BLSTM)

We use the baseline data of 6,072 VA candidates with 3,023 confirmed VAs to train and test a BLSTM on whole sentences (Schuster and Paliwal, 1997; Graves and Schmidhuber, 2005) using 5-fold cross validation. The BLSTM classifies sentences on whether they contain a VA expression or not. We represent each word from a sentence candidate with a pre-trained word embedding. We use word embeddings from GloVe which consist of 300-dimensional vectors that were trained on a Google News corpus (Pennington et al., 2014). We implement the BLSTM in Keras (Chollet et al., 2015) with Tensorflow backend, using Adam Optimizer and default hyperparameters (epochs=80, batch size=128, hidden units=300, dropout=0.25). We expect that further improvements may be possible if a fine-grained hyperparameter search was to be conducted.

5 Evaluation and Results

Determining an overall *recall* of our approaches on the New York Times corpus is unfeasible as it contains 1.8 million articles. Thus, we used a random sample of 105 articles (for each year 5). However, we found only one VA: "If Nike is the Chicago Bulls of the athletic shoe market, retailers are holding their breath for a strong underdog player to emerge." (NYT 1997/06/07/0935205). Due to our focus on individuals, it is no VA that should be part of our data set. Thus, we can only determine *precision* and *recall* of our new approaches based on the baseline data set described in Section 3.

5.1 Difficulty of the Task

Identifying VAs is a non-trivial task even for humans. Thus, we used *expert knowledge* to evaluate candidates for the ground truth data (see Section 3).

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%

Table 2: Performance of the three proposed approaches in comparison with the baseline.

In addition, we leveraged *crowdsourcing* to check whether we could use (untrained) people to evaluate VA candidates. Therefore, workers on the crowdsourcing platform Figure Eight had to check whether a sentence contained a VA.[10]

Crowdsourcing resulted in 600 judgments for 200 randomly selected candidates of our baseline data set with an inter-annotator agreement calculated by Cohen's Kappa of 0.72 between expert and workers. Re-checking the 28 disagreements showed that the expert judged all correctly, which results in a crowdsourcing accuracy of 86%. This shows how difficult it is to identify VAs.

5.2 Baseline VA Candidates

We cannot determine recall for our approaches, instead we compute precision, recall and f-score based on the baseline data set (see Section 3), which is shown in Table 2 as well as precision of the baseline data set. The automated approaches using Wikidata and named entity recognition can boost the precision significantly with a moderate loss in recall. The BLSTM approach performs best. Although the loss in recall is higher than with the WD approach, the precision reaches almost 87% and is thus raised to a new level.

5.3 Non-Baseline VA Candidates

The results of our approaches are not limited to the baseline data set. The WD approach generated 955 new VA candidates, the NER approach 4,399. The human annotators evaluated 100 randomly selected VA candidates of each set, which resulted in a precision of 17% (WD) and 30% (NER). We also predicted the labels of these random samples with the trained BLSTM which resulted in 83% and

---

75% precision on the WD approach and the NER approach, respectively. This shows that the BLSTM does not only have potential to improve the results of the baseline approach but also of the WD and NER approaches.

5.4 Error Analysis

About 13% of the WD approach errors (based on baseline) were false negatives, for example, "an Edison of magic". Here, "Thomas Edison" would have been the right source, but as the entity "Thomas Edison" has no alias "Edison", the most popular entity having 'Edison' as name or alias is a town in the United States.[11] 87% were false positives like "the Michelangelo of the Sistine Chapel" where the person name was not used as a VA source but stands for a specific work period of the artist. In the sample set that we selected randomly from the WD approach candidates which did not appear in the baseline (see Section 5.3), we detected false positives like "an Air of Mystery" as "Air" is an alias of Michael Jordan, who has more sitelinks than the word "air" (116 to 113).

The errors from the NER approach (based on baseline) had around 36% false negatives like "the Marco Polo of baseball" where the source was not detected as a person by the named entity tagger. 64% were false positives like "the Dave Brown of old". The NER approach sample set from Section 5.3 contained new false positive candidates because awards ("the Harriet H. Jonas Award of"), institutions ("the O. K. Harris Gallery of"), or titles ("the Episcopal Bishop of") were falsely tagged as persons.

5.5 Discussion

Since we presented the first approaches for the extraction of VAs that work completely automatically, our work has some limitations that we discuss in the following. The restriction to one type of source can be solved easily, for instance, by choosing not only 'humans' but proper nouns in general, as well as allowing further syntactic variations of the source phrase, for example, 'the ADJECTIVE ENTITY' ("the new Michael Jordan") or 'the ADJECTIVE/NOUN sort/version/equivalent of' ("the Georgian equivalent of") but (probably) at the expense of precision.

Beyond extending recall by tackling further syntactic variations and allowing non-human sources,

we currently aim to extract the modifier and target of VAs.

Another goal is to train a neural network with a larger and better balanced training set to use the model to study a larger corpus. Alternatively, a pretrained model could be used and fine-tuned with our labeled data to improve results.

Besides above-mentioned limitations which we plan to address in future work, we also want to report on approaches that have not lead to any improvements. For example, we tested measuring the similarity between source and modifier using word embeddings (with the assumption that a low similarity should indicate a VA). Unfortunately, there are too many neutral modifiers like "his time" or subgenres of the source like "the Tiger Woods of micro golf" where this idea does not work. Furthermore, removing candidates whose source was contained in WordNet (Fellbaum, 1998) and not labeled as 'person' did not work, since "wife", "chancellor", and so forth were labeled as persons in WordNet.

6 Conclusions

In this paper, we presented approaches for the first fully automated extraction of the stylistic device *Vossian Antonomasia*.

In addition, we were able to create the largest known collection of VAs by looking into 21 years of the New York Times. The data set contains the annotated list of 6,072 VA candidates, each with article id, date, URL, marked source and modifier, Wikidata id, and class.

The best approach, using a BLSTM, reached a precision of 86.9% and a recall of 85.3%.

References

Kateryna Akhtyrska. 2014. *Linguistic Expression of Irony in Social Media*. Ph.D. thesis, Universidade do Algarve.

Angelika Bergien. 2013. Names as frames in current-day media discourse. In *Name and Naming. Proceedings of the second international conference on onomastics*, pages 19–27, Cluj-Napoca. Editura Mega.

François Chollet et al. 2015. Keras. https://keras.io.

Ben Cohen, Geoff Foster, Matthew Oshinsky, Jon Keegan, and Tom McGinty. 2015. The Michael Jordan of .... Website.

James Fallows. 2014. The Steve Jobs of Beer. *The Atlantic*, (November 2014).

---

[11] https://www.wikidata.org/wiki/Q746801

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Frank Fischer and Robert Jäschke. 2019. 'The Michael Jordan of greatness'—Extracting Vossian antonomasia from two decades of The New York Times, 1987–2007. *Digital Scholarship in the Humanities*, (fqy087).

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural Metaphor Detection in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 607–613. Association for Computational Linguistics.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5):602–610.

Kenneth Holmqvist and Jarosław Płuciennik. 2010. Princess Antonomasia and the Truth: Two Types of Metonymic Relations. In Armin Burkhardt and Brigitte Nerlich, editors, *Tropical Truth(s). The Epistemology of Metaphor and Other Tropes*, pages 373–381. De Gruyter, Berlin/New York.

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word Embedding and WordNet Based Metaphor Identification and Interpretation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1222–1231. Association for Computational Linguistics.

Malvina Nissim and Katja Markert. 2003. Syntactic features and word similarity for supervised metonymy resolution. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 56–63. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543. Association for Computational Linguistics.

Evan Sandhaus. 2008. The New York Times Annotated Corpus LDC2008T19. DVD, Linguistic Data Consortium, Philadelphia.

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *Transactions on Signal Processing*, 45(11):2673–2681.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor Detection with Cross-Lingual Model Transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 248–258. Association for Computational Linguistics.