“Who is the Madonna of Italian-American Literature?”: Extracting and Analyzing Target Entities of Vossian Antonomasia

Michel Schwab[1], **Robert Jäschke**[1,2] **and Frank Fischer**[3]
[1]Humboldt-Universität zu Berlin, Germany
[2]L3S Research Center, Hannover, Germany
[3]Freie Universität Berlin, Germany
{michel.schwab,robert.jaeschke}@hu-berlin.de
fr.fischer@fu-berlin.de

Abstract

In this paper, we present approaches for the automated extraction and disambiguation of a part of the stylistic device Vossian Antonomasia (VA), namely the target entity described by the expression. We model the problem as a coreference resolution and a question answering task and also combine both. To tackle the tasks at hand, we utilize state-of-the-art models in these areas. In addition, we visualize the connection between source and target entities of VA in a web demo to provide a deeper understanding of their mutual relationship.

1 Introduction

Vossian Antonomasia (or *VA* for short) is a popular stylistic device used to describe an entity by refering to another entity, typically in a witty and resourceful way. Structure-wise, a VA expression consists of three parts: target (trg), source (src), and modifier (mod). The combination of source and modifier is used to describe the target. Take, for instance, the sentence “It is the Madonna of Italian-American literature in that it shows the transition from the Italian immigrant to American citizen like no other book of its genre.” (NYT 1991/08/07/1128838).[1] The author uses “Madonna”, the popular American singer, as source and transfers a set of characteristics of Madonna to the target, Helen Barolini’s novel “Umbertina”. The modifier “Italian-American literature” projects these characteristics onto the target.

In general, the source consists of a universally known, famous named entity, from which one or more typical traits or characteristics are to be invoked. The target, on the other hand, does not necessarily have to be a named entity (e.g., “Rally$_{\texttt{trg}}$ car$_{\texttt{trg}}$ racing$_{\texttt{trg}}$ is the David$_{\texttt{src}}$ Hasselhoff$_{\texttt{src}}$ of motor$_{\texttt{mod}}$ sports$_{\texttt{mod}}$” (NYT 2006/08/02/1780256)). It is also possible that no specific target is meant, in which case the VA expression would be hypothetical, for instance, “We’re waiting for the Raffi$_{\texttt{src}}$ of our$_{\texttt{mod}}$ industry$_{\texttt{mod}}$.” (NYT 1989/06/11/0257799), or there is a target, but it is not explicitly mentioned in the article content.

The task of extracting VA expressions focusing on source and modifier has already been covered (Fischer and Jäschke, 2019; Schwab et al., 2019, 2022). The models are also able to identify a reference of the target inside a sentence where source and modifier appear, but in most cases, these references are in the form of pronouns or other mentions of the entity that cannot be linked to a knowledge base for disambiguation. To the best of our knowledge, there exists no method for extracting the target entity from texts. As the phrase consisting of source and modifier, such as “the Madonna of Italian-American literature”, is a specific mention of the target entity (“Umbertina”), it should appear in the reference chain of the target when using coreference resolution. However, as we will later show, coreference resolution often fails to detect these complex mentions.

For a deeper understanding of VA expressions and their meaning, the target entity is an essential part. We need to identify it to comprehend the transferred characteristics. After its identification, we can compute VA chains where the source of one VA expression is the target of another to track the transfer of characteristics across multiple entities and also analyze the assignments of characteristics to entities. Thus, in this paper, we tackle two tasks:

Target extraction: The automatic extraction of the full name of the target entity inside a text.

Target Linking: Disambiguation and linking to Wikidata.

In addition, we visualize the results in a web demo for exploration and visualization.[2]

[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]https://vossanto.weltliteratur.net/sighum2023/

Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL2023), pages 110–115
May 5, 2023 ©2023 Association for Computational Linguistics

Our annotated data (Schwab et al., 2023) and code are freely available.[2]

2 Related work

The detection and extraction of VA has recently been worked on. While Jäschke et al. (2017); Fischer and Jäschke (2019) used semi-automated approaches to detect VA expressions, Schwab et al. (2019) developed the first automated approach for the detection of VA expressions on the sentence-level. They developed a finer extraction approach on the word-level (Schwab et al., 2022). In particular, they employed pre-trained contextual language models, for instance, BERT (Devlin et al., 2019), and fine-tuned them on an annotated dataset modeling the problem as a sequence tagging task. However, all models lack the ability to identify the target entity within the article.

The automated detection and extraction of other stylistic devices, such as metaphors, has been covered widely. The extraction of VA expressions consisting of source and modifier is closely related to metaphor detection. However, our focus is on the extraction of the target entity, so we do not consider such related work.

3 Annotation and Methods

3.1 Dataset and Annotation

We use the dataset from Schwab et al. (2022), which is an annotated VA dataset on the word-level. The dataset consists of 5,995 sentences, of which 3,066 contain VA expressions and 2,929 do not. In this paper, we focus only on sentences containing VA expressions. The dataset originally emerged from Schwab et al. (2019) who used nine syntactic patterns to identify VA candidates focusing on the syntax around the source entity. Those candidates were extracted from The New York Times Annotated Corpus (Sandhaus, 2008) which contains more than 1.8 million newspaper articles from the years 1987 to 2007. Thus, the syntax around the source of each VA expression in the dataset consist of one of the nine following variations: "a/an/the SOURCE of/for/among MODIFIER" (e.g., "the Madonna$_{\mathtt{src}}$of Italian-American$_{\mathtt{mod}}$literature$_{\mathtt{mod}}$"), which we refer to as "VA phrase" in the sequel.

The target annotation in this dataset is limited to a reference within the sentence where the VA phrase appears, which mostly does not include the target's name but pronouns (e.g., "she"), other denotations (e.g., "the president"), or which does not exist at all.

Sentences in the dataset may include multiple VA expressions. In order to separate them, we created copies of such sentences for each VA expression with just one annotation resulting in 3,115 sentences. Two trained students annotated all target names inside each NYT article containing a VA expression from the dataset by Schwab et al. (2022). Specifically, the annotators took a closer look at the article in which the sentence occurred and extracted the name of the entity to which the VA expression referred. In other words, they conducted coreference resolution modeling the VA expression as one reference of the target. In addition, they linked the marked entity to the corresponding Wikidata entity, if available, and extracted the Wikidata ID. This resulted in an inter-annotator agreement calculated by Cohen's Kappa of 0.96 (annotation) and 1.0 (linking), measured on a sample of 500 randomly selected sentences in the dataset. Disagreements were discussed and then re-annotated. In 2,853 (91.6%) of the cases, there existed a target name. In all other examples, there was no mention of the target and therefore we omit these cases for our study. In 2,354 (75.6%) of the cases, the annotators were able to link the name to the corresponding Wikidata entity. The absence of Wikidata entries for the remaining target entities could be due to a lack of prominence or relevance.

3.2 Coreference Resolution (COREF)

In our annotated dataset, there exist two references to the target entity: The first is the VA phrase itself (e.g., "the Madonna of Italian-American Literature") as explained previously. The second is the mention of the target inside the same sentence as the VA phrase. Schwab et al. (2022) annotated the mention (e.g., "It", "Mr. Woods"), if it existed and their models are able to identify this reference together with the source and modifier within a sentence. Both expressions should be part of the reference chain of the target entity.

Thus, the obvious choice to tackle the problem of extracting the target entity is coreference resolution as it is already well-studied. Because of this, we will use the coreference resolution model as baseline.

Modern coreference resolution systems show strong results but lack the size of the input document. To tackle this problem, Beltagy et al. (2020) introduced Longformer, pre-trained language models that are able to handle input documents with up to 4,096 tokens. Toshniwal et al. (2021) picked up this idea and used the Longformer model as a base for their coreference resolution model which showed state-of-the-art results for a variety of datasets. Therefore, we will use this model to perform coreference resolution on the entire article text and create two baselines. In the first one, we select the reference chain that includes the VA phrase, whereas in the second, we choose the chain of the annotated target mention. As our task is to find the full target name rather than the reference chain including the name, we need to choose a mention from the chain as output. To do this, we utilize a named entity tagger, specifically the NER model from Akbik et al. (2018). The tagger identifies all named entities that appear in the reference chain. We then select the first named entity that emerges in the chain, based on the assumption that authors usually introduce named entities with their complete names in article texts.

$\mathbf{LF_{\{p,t\}}}$: We use the joint model from Toshniwal et al. (2021), which is fine-tuned for coreference resolution over a mixture of datasets (OntoNotes, LitBank, and PreCo). LF_p refers to the model that focuses on the reference chain of the VA phrase, LF_t concentrates on the chain that includes the target mention.

3.3 Question Answering (QA)

Coreference resolution is one way to tackle the problem. However, as we do not look for the complete reference chain but only for the name of the entity, coreference resolution is not needed after all. Another way to solve the problem is to re-formulate the task as an extractive question answering problem by using the advantage of the annotated VA phrase within the sentence. Since the VA phrases are syntactically similar (see Sec. 3.1), we use them to formulate the query: "Who is the/a/an SOURCE of/for/among MODIFIER?". For the task of extractive question answering, we need to give the model a context text to extract the answer from.

In one scenario, we use the complete article content the VA expression appears in as context. In another scenario, we only use the content before the sentence that includes the VA expression together with the sentence itself and the subsequent 200 characters. In a preliminary analysis, we found that the target entity is typically mentioned earlier in the article, thus the noise of the rest of the article may decrease the performance of the model. Still, in some cases, the target entity is mentioned shortly after the VA expression. Thus, we include 200 characters after the sentence with the VA phrase which covers more than 98% of all cases.

Similar to coreference resolution, QA is a widely studied task. Therefore, instead of training a new model, we use a state-of-the-art fine-tuned language model, namely the one from Clark et al. (2020). The problem of the length of our documents (articles) is tackled by a sliding window approach.

$\mathbf{ELE_{\{c,s\}}}$: We employ the ELECTRA large model that is fine-tuned on the SQuAD2.0 dataset using both context scenarios. ELE_c refers to the complete context, ELE_s to the short context scenario.

3.4 Hybrid Approach

In a third approach, we combine both methods, using QA first and coreference resolution on top of it. In some cases, the QA models return an answer which is not the full target name but only another reference of the target entity, for instance, "Mrs. Merkel" instead of "Angela Merkel". This is not a correct output for our task. Thus, we apply coreference resolution on the QA output to get the entire reference chain. As in the baselines, we identify all named entities in the selected chain. Then, we leverage the QA output and choose the longest named entity (in terms of characters) that shares at least one word with the QA output. For instance, if the QA output is "Mr. McCaw" and the named entities in the reference chain include "McCaw", "Craig O. McCaw" and "Mr. McCaw" (in this order), we select "Craig O. McCaw" as the output. This is because it shares a word ("McCaw") with the QA output and is longer than the other candidates that share at least one word ("McCaw", "Mr. McCaw"). In the baseline scenarios, "McCaw" would have been selected. This heuristic approach surpasses the performance of multiple fuzzy string matching algorithms, such as the Levenshtein distance (Levenshtein et al., 1966) and Jaro-Winkler similarity (Winkler, 1999).

ELE+LF: We concatenate both methods using ELE and LF.

3.5 Entity Linking (EL)

In a second step, we aim to disambiguate the entities found with the previous methods and link them to their corresponding Wikidata entries. For this, we employ GENRE (De Cao et al., 2021), a state-of-the-art entity linking approach. GENRE is a sequence-to-sequence model that is based on a fine-tuned BART architecture (Lewis et al., 2020) which links the given input entity to a Wikipedia entity using the surrounding context. In particular, it is a generation model that generates the output using constrained beam search. As each Wikipedia entry has a unique Wikidata entry, we can get the Wikidata ID of the Wikipedia entry using the MediaWiki Action API.[3] If the output of the extraction method is a reference chain, we conduct entity linking on each mention in the chain separately, skipping all pronouns. For the baselines, we select the prediction that appeared most frequently. For $\mathrm{ELE_s}$+LF, we choose the prediction that shares the largest word overlap with the QA output. If predictions are equally frequent or share an equal number of overlapping tokens, we choose the prediction that is closest to the beginning of the text. Consider, for instance, the reference chain consisting of "Lomax", "Alan Lomax, the musicologist who evangelized folk music for most of the 20th century", "the Johnny Appleseed of folk revivalists", "Alan Lomax" and "Lomax" and the QA output "Alan Lomax". The EL predictions for the chain are "Alan Lomax", "Alan Lomax", "Johnny Appleseed", "Alan Lomax" and "John Lomax". The highest share of words between the QA output and the predictions is "Alan Lomax' which we take as output.

GEN: We use the entity disambiguation GENRE model that is pre-trained on the BLINK dataset and fine-tuned on the AIDA CoNLL-YAGO dataset.

4 Results

4.1 Evaluation

We use two different evaluation metrics in order to evaluate the target extraction models. The first metrics, namely precision, recall and F_1, are based on the overlapping tokens of prediction and ground truth. The second metric, exact match (em), measures the percentage of predictions that fully match the ground truth.

Additionally, we evaluate the entity linking model. For that, we use InKB micro precision (mp). Micro precision describes the share of correctly linked entities and InKB, which is introduced in Röder et al. (2018), means that we only consider entities that have a valid Wikidata entry.

[3] https://www.mediawiki.org/wiki/API:Main_page

	Extraction				Linking
model	prec	rec	f1	em	mp
$\mathrm{LF_p}$	.29	.28	.29	.25	.21
$\mathrm{LF_t}$	.58	.54	.56	.47	.46
$\mathrm{ELE_c}$	.66	.62	.64	.52	.55
$\mathrm{ELE_s}$	.74	.71	.72	.61	.62
$\mathrm{ELE_s}$+LF	.78	.77	.78	.71	.64

Table 1: Performance of the three proposed approaches in comparison with the baselines ($\mathrm{LF_{\{p,t\}}}$).

The results, presented in Table 1, demonstrate that the baselines are unable to solve this task effectively. In particular, $\mathrm{LF_p}$ shows that coreference resolution on the VA phrase does not work as expected. $\mathrm{ELE_c}$ has a significant gap in comparison to $\mathrm{ELE_s}$, where we used our trick of truncating the context, across all metrics. The combined model, $\mathrm{ELE_s}$+LF, outperforms all other models, especially in the em score by, a large margin.

While GEN has an upper limit of .84 (mp) on the annotated target names, a score of .64 is not necessarily poor. However, it does suggest that there is room for improvement and underscores the overall complexity of entity linking in general.

4.2 Error Analysis

$\mathrm{LF_{\{p,t\}}}$: Most errors in the baselines occured because the selected reference chains did not include the full target name. In particular, only in 887 (31.0%) and 1,687 (58.4%) of the cases for $\mathrm{LF_p}$ and $\mathrm{LF_t}$, respectively, the reference chain include the full target name. That shows the difficulty for the baseline models to achieve a better result. Additionally, finding the correct mention in the chain was challenging. Only in around 80% of all instances, the correct mention was chosen when the correct reference chain had been select before. This is because the first named entity in the reference chain is not always the full target name, e.g., when the VA phrase appeared in the reference chain before the full target name in the article.

$\mathrm{ELE_s}$+LF: In our best approach, the correct reference chain was found in 72%. Choosing the correct mention in the reference chain worked well. Only in around 1% of all instances, our approach did not select the correct mention. Interestingly, $\mathrm{ELE_s}$ provided the correct answer in 1746 (61%) of the cases. In around 20% of all instances, it found a mention of the target entity where the correct name was in-

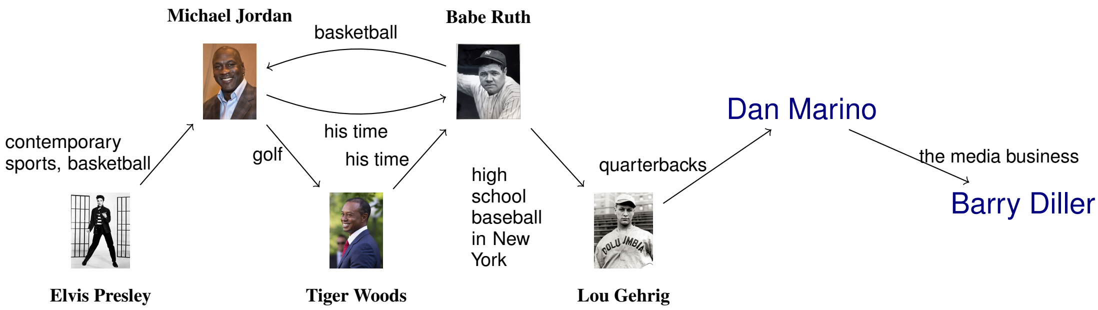

Figure 1: The longest VA chain in the dataset.

cluded, e.g. "Mr./Ms./Mrs./Dr. surname" or first names in direct speech, i.e. another reference of the target entity. Most of the other false predictions consisted of incorrect chosen entities.

In addition, we tested two hypotheses regarding the impact on the performance (em score) of the model using the Point-Biserial Correlation Coefficient as the em score is a binary label. The first hypothesis is whether there exists a negative correlation between the character-wise distance of the full target name and the VA phrase in the article text and the performance. The results show a weak negative correlation between the two variables (-0.131), and, with a p-value of 0.00, it is statistically significant. This result can be interpreted that the distance does have a slight negative effect on the model's performance. The second hypothesis examines whether the size of the reference chain that includes the full target name is important for the model's performance. We assumed that it might be easier for the model to predict the correct cluster if the target entity is an important figure in the article and thus, which normally results in a larger cluster size. However, the r-value of 0.035 indicates almost no correlation between the two variables and as the p-value is 0.088, the assumption is not statistically significant and should be withdrawn.

4.3 Application Scenario

From the point of view of stylistics, VA is a powerful device because it can not only "spice up" a text, but can also set a decisive accent through its often surprising suggestiveness, which is why it is also well suited for headings or subheadings. So far, due to a lack of available data, the phenomenon has not been analyzed on a large scale, specifically, the relationships between source and target entities. In this paper, we lay the groundwork for such analysis, which enables the exploration of the transfer of characteristics between different entities. To accomplish this, we visualize the results in a web demo.[4] In particular, we model the source and target entities as nodes in a network and connect them with edges when they co-occur in a VA expression. The web demo displays the annotated dataset. It helps to explore chains between entities and can provide new insights in the use of VA and the choice of entities, see Figure 1.

5 Conclusion

We have shown that the extraction of the target name and its linking is not a trivial task, and that state-of-the-art coreference resolution models, which should cover this task, do not perform as well as they do on common datasets in their domain. However, our idea of modeling the problem as a question-answering task by employing the annotation of source and modifier shows better results and the concatenation of both models, first ELE_s followed by LF shows promising results. Notably, re-formulating the VA expression into a QA problem works on all syntactic forms of VA expressions and is not limited to the syntax in the dataset. These findings also show that the annotated VA dataset that emerged from the target annotations, even though it is a specific device, can be used as an out-of-domain evaluation dataset for QA and COREF models in general.

With the disambiguation of the target entity we are able to deepen the understanding of VA. The ability to track VA chains is a completely new field that can lead to many interesting insights into the function and use of VA, for example, regarding the transfer of characteristics. Our web demo greatly simplifies the exploration of connections between source and target entities.

[4] https://vossanto.weltliteratur.net/sighum2023/graph.html

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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*, 35.

Robert Jäschke, Jannik Strötgen, Elena Krotova, and Frank Fischer. 2017. “Der Helmut Kohl unter den Brotaufstrichen”. Zur Extraktion Vossianischer Antonomasien aus großen Zeitungskorpora. In *Proceedings of the DHd 2017*, DHd '17, pages 120–124. Digital Humanities im deutschsprachigen Raum.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Michael Röder, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. 2018. Gerbil–benchmarking named entity recognition and linking consistently. *Semantic Web*, 9(5):605–625.

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

Michel Schwab, Robert Jäschke, and Frank Fischer. 2022. "The Rodney Dangerfield of Stylistic Devices": End-to-end detection and extraction of vossian antonomasia using neural networks. *Frontiers in artificial intelligence*, 5.

Michel Schwab, Robert Jäschke, and Frank Fischer. 2023. Annotated vossian antonomasia dataset.

Michel Schwab, Robert Jäschke, Frank Fischer, and Jannik Strötgen. 2019. 'A Buster Keaton of Linguistics': First automated approaches for the extraction of Vossian Antonomasia. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, EMNLP '19, pages 6239–6244. Association for Computational Linguistics.

Shubham Toshniwal, Patrick Xia, Sam Wiseman, Karen Livescu, and Kevin Gimpel. 2021. On generalization in coreference resolution. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 111–120, Punta Cana, Dominican Republic. Association for Computational Linguistics.

William E Winkler. 1999. The state of record linkage and current research problems. *Statistical Research Division, US Bureau of the Census, Washington, DC*.