# "Der Frank Sinatra der Wettervorhersage": Cross-Lingual Vossian Antonomasia Extraction

**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

We present a cross-lingual approach for the extraction of Vossian Antonomasia, a stylistic device especially popular in newspaper articles. We evaluate a zero-shot transfer learning approach and two approaches that use machine-translated training and test data. We show that our proposed models achieve strong results on all test datasets in the target language. As annotated data is sparse, especially in the target language, we generate additional test data to evaluate our models and conclude with a robustness study on real-world data.

## 1 Introduction

Automatic detection and extraction of stylistic devices is an important task for understanding natural language, especially for understanding figurative language. However, most research focuses on English corpora, since labeled data is, to our knowledge, only available in English. Often those approaches cannot be applied to other languages for various reasons, for example, the lack of labeled data, syntactic variations, or semantic differences. In this paper, we study the cross-lingual extraction of the stylistic device Vossian Antonomasia (VA).

VA is a specific kind of antonomasia closely related to metonymies and metaphors. In short, it is used to describe an entity (target) by mentioning another entity (source) and a context (modifier) that refers to the target. Usually, the source entity is famous and well-known to the reader in order to understand the author's intentions. Particularly, a set of characteristics of the source is used implicitly to describe the target. The modifier shifts the source's characteristics to the target's environment. The title of this paper may serve as an example. In a German newspaper article (elm, 2013) Elmar Gunsch (nicknamed "Die Stimme"[1]) is called the "Frank Sinatra of weather forecasting". It is also explained why: "Elmar Gunsch's sonorous bass brought glamor to ill-humored German television in the seventies." The popular American singer and actor Frank Sinatra serves as the source of this VA. Through the formulation, some of Sinatra's characteristics (voice, entertainment qualities, and popularity) are transferred to the target, Elmar Gunsch. The dependent genitive construction ("of weather forecasting") represents the modifier.

The automated detection and extraction of VA is a non-trivial task as the syntax can be ambiguous. Take, for instance, "the Galileo of welfare reform" (Roberts, 1992) vs. "the Galileo of the Fantasy line" (gal, 1990). While the first example is a VA expression referring to a senator, the latter is the name of a cruise from a cruise line company. This is also one of the reasons why rule-based approaches fail, as seen in Schwab et al. (2019) where a trained neural network outperforms the rule-based approaches. While the extraction of VA from English corpora has already been studied Schwab et al. (2019, 2022), there are no studies on automated approaches to find VA in other languages, yet. This is the starting point for this paper: we study the automated extraction of VA in another language – German.

As the lack of annotated data is one of the main problems for the study in different languages, we will use three different sequence tagging approaches: The first is based on zero-shot transfer learning, the second and third on machine translation and word alignments of the annotated data. All models employ pre-trained language models because they achieved the best results on English VA extraction, see Schwab et al. (2022).

Despite the sparse occurrence of VA in text corpora, it can assist several NLP applications. Co-reference resolution can be supported as the source entity should not be a single co-reference chain but together with the modifier part of the target chain. The generation of fruitful and interesting

[1]"The Voice", our translation

| |
|---|
| **Lanz**, der OBAMA des *deutschen Fernsehens*<br>**Lanz**, the OBAMA of *German television* |
| Ein *spanischer* LIONEL JOSPIN müsse her.<br>A *Spanish* LIONEL JOSPIN was needed. |
| **Statine**, der BENTLEY unter *den Kardioprotektiva*<br>**Statins**, the BENTLEY of *cardioprotectants*. |

Table 1: Three examples of German VA expressions together with their translation and alignment.

text is another reason, especially the generation of headlines. It also assigns attributes to entities and connects entities together which can lead to interesting question-answering tasks. In general, it is a use case for similar cases where there is a lack of large annotated datasets and can therefore show ways to tackle similar downstream tasks that are not as rich as common NLP tasks as named entity recognition.

In the following, we want to answer whether 1) our models can compete with mono-lingual approaches, 2) machine translation based models can compete against new zero-shot models, and 3) our models are robust against real-world data. Our code and data are freely available.[2]

## 2 Related Work

The automated extraction and detection of VA has not been studied deeply. There exist rule-based approaches (Fischer and Jäschke, 2019; Schwab et al., 2019), the latter also trained a neural network based on non-contextualized word-embedding and bi-directional LSTM layers to classify sentences into whether they include VA expressions or not. Schwab et al. (2022) constructed neural network models from scratch but their best models are based on pre-trained language models, BERT (Devlin et al., 2018). In addition to a binary sentence classifier, the authors tackled the problem of detecting all chunks of a VA expression inside a sentence. They created an annotated dataset and showed that their models are robust on real-world data. They also showed that adding unlabeled random sentences to their training data as negative instances improved the model. Adding such sentences helped to diversify the syntactic variations of negative instances and also generated a class imbalance that is closer to real-world data without the cost of annotation.

Cross-lingual approaches for detecting similar stylistic devices, for instance, cross-lingual metaphor detection, have also not been studied deeply: Tsvetkov et al. (2014) used lexical semantic features and word embeddings (ENG-{ESP, FAS, RUS}), whereas Vída (2020) studied the problem of cross-lingual metaphor paraphrase detection (ENG-CZE). They used machine translation, but also multilingual word embeddings, MUSE (Lample et al., 2018). Aghazadeh et al. (2022) presented models that employ the layers of pre-trained language models in a zero-shot probing scenario.

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

| training data | | test data | |
|---|---|---|---|
| before | after | before | after |
| the | of | die/der/das | der/des |
| a/an | for | ein/e | von/vom |
| | among | diese/r/m | unter |
| | | Art | aus |
| | | als | für |
| | | den/dem | - |

Table 2: Most common syntactic variations around the source in the training and test data. The column 'before' shows the words appearing before, 'after' the words following the source in the sentence. '-' signals that there is no boundary word after the source but either the end of the sentence or the modifier.

## 3 Datasets

In this section, we describe the datasets and the annotation process. The training data is in English whereas the test data consist of German texts. Comparing the diversity of the datasets, we can clearly see that the test data is richer in terms of the syntactic variations of VA, see Table 2. The table shows the most frequently used boundary words around the source. Whereas the training data has a limited collection of boundary words, the test data shows a great diversity. This is because of the manual collection of one of the test data which did not follow any syntactic restrictions like the training data.

### 3.1 Training Data

**NYT-0:** The dataset originally emerged from a semi-automatic rule-based approach from Schwab et al. (2019) on The New York Times Annotated Corpus (Sandhaus, 2008). Schwab et al. (2022) expanded the dataset and tagged all VA chunks inside a sentence on the word level. We updated the target annotations to improve consistency (cf. Section 3.3) and corrected mislabeled annotations. The dataset contains 3,065 sentences with VA expressions (which we call *positive* in the sequel) and 2,930 without (*negative*). All VA expressions in this dataset follow a specific syntax, that is, the/a/an SOURCE of/for/among, see Table 2.

**NYT-50:** Like Schwab et al. (2022) we add 50,000

random sentences from the New York Times Corpus (Sandhaus, 2008) to the NYT-0 dataset to diversify the negative instances in the dataset and to tackle the biased share of positive instances in the NYT-0 dataset ($\approx 50\%$).

### 3.2 Test Data

**UMBL:** This VA collection[3] was manually gathered from German newspaper articles but also in radio, TV, or videos between 2009 and 2014. It only contains sources and modifiers of VA expressions, but not the targets and the original sentences. We tried to collect the sentences from the original articles, but could not use all instances as some of the articles the expression appeared in were behind paywalls or did not exist anymore. Out of 470 positive instances, we could get full information in 362 cases which we annotated as explained in Section 3.3. The collection is not based on any syntactic rules, thus, the VA expressions contain broader syntactic variations compared to the NYT dataset, as can be seen in Table 2. Also, in this dataset the modifier may appear before the source, see Table 1, Ex. 2 which does not appear in the NYT dataset.

**ZEIT:** A dataset where Jäschke et al. (2017) used a complex rule-based approach to collect VA instances from German weekly "Die Zeit" (covering 1995 to 2011). In total, they found 1,456 candidate sentences of which 224 are positive. The 224 instances together with the sentence they appear in are publicly available and fit as a test dataset. Source and target were already annotated which left us to annotate the modifier.

**Generation of negative data:** As both datasets only contain positive instances, we generated additional samples that consist of negative instances to evaluate our model accurately. We use the sparseness of the phenomenon to create one random sample and two focused samples that contain instances similar to those in the training data in terms of syntax or the choice of entities to make sure that none of these reasons are biasing our models. Each of those samples includes 3,000 sentences.

**NEG1:** The dataset contains sentences which include phrases that are syntactically similar to the VA expressions in the NYT-0 dataset regarding source and modifier. That is, the modifier appears behind the source, mostly separated by a delimiter word (e.g., "der/des" - "of"), see Ex. 1, Table 1. So, we extracted all Wikidata entities that have a German label which contains one of the delimiter words between two other words, for example, "Königin von England". We then took a sentence from the corresponding German Wikipedia web page that included the label and use this sentence as an instance in our test dataset.

**NEG2:** Most of the source entities in all three datasets are humans. Thus, we want to analyze whether the choice of entities in test instances has an impact on the prediction. We use a corpus of a German newspaper, "taz, die tageszeitung" (TAZ) (200, 2005) to create the sample. This corpus contains more than one million German news articles from 1985 to 2005. For each of the 1,691 distinct source entities in the NYT-0, UMBL and ZEIT datasets, we extracted three random sentences in the TAZ corpus that include the entity's name. In total, we could extract 3,940 sentences and again we used a sample of 3,000 sentences. Due to different reasons, for example, different spellings, we could not find sentences for all entities.

**NEG3:** is generated by a random selection of sentences from the TAZ corpus mentioned above.

### 3.3 Annotation Process

In the UMBL and ZEIT corpus, we annotated different chunks, whereas in the NYT-0 dataset, we only updated the targets due to consistency, see below. For NEG-1, NEG-2 and NEG-3, all found VA expressions were replaced by negative instances to keep those samples consist without any VA expressions.

We follow the IOB annotation scheme from Schwab et al. (2022) with one exception: For target annotations, we annotate the whole noun phrase instead of the entity only (leaving out relative clauses due to length) as this has not been annotated consistently before. The annotations of NEG1, NEG2, and NEG3 were conducted by one trained annotator. The annotator found 43 VA expressions in NEG-2, 3 in NEG-3 and none in NEG-1. Those were replaced by negative sentences following the generation process for each sample to have consistent negative samples. Additionally, two other trained annotators annotated 100 random instances of each sample for the quality assessment of the annotations which resulted in a full agreement of all instances except one which was discussed and annotated accordingly.

[3] https://umblaetterer.de/datenzentrum/vossianische-antonomasien.html

## 4 Methods

We model the problem as a sequence tagging task (Schwab et al., 2022): For each word $w_i$ of sentence $S$ predict a tag $t_i$ which indicates whether $w_i$ is part of target, source, or modifier, or not a VA part at all.

We study a zero-shot cross-lingual transfer scenario as well as models that use machine-translated test or training data. The fine-tuning step in all three methods is conducted by adding a linear layer on top of the pre-trained model architecture that outputs a tag for each token of the input.

**0shot:** Zero-shot approaches on multilingual language models have recently shown great advances. They are often used in cases like ours, that is, there exists no or only few annotated data in the target language. In short, a language model is pre-trained on a multilingual corpus and then fine-tuned only with the annotated data from one language (source language). Without seeing any annotated data from the target language, it has been shown that those fine-tuned models are able to transfer their learning to languages that they have been pre-trained with, see Conneau et al. (2020). We fine-tune the XLM-RoBERTa model (Conneau et al., 2020) with the NYT training data and evaluate it on the test data.

**de2en:** We translate the German test data to English using machine translation, in particular the FAIRSEQ toolkit (Ott et al., 2019). We align the original translated sentence pairs with a word alignment tool (Jalili Sabet et al., 2020) and project the corresponding tags to the translated sentences.[4] Then, we fine-tune a BERT model (Devlin et al., 2018) with the NYT datasets and evaluate it on the translated and aligned test data.

**en2de:** We use the architecture as in de2en but the other way round: We translate the training data to German using FAIRSEQ and project the tags with the word aligner.[4] Then, we fine-tune a German pre-trained language model, DBMDZ BERT,[5] with the translated data and evaluate it on the test data.

## 5 Experiments

### 5.1 Experimental setup

Hyperparameter optimization is applied on epoch ($e$), learning rate ($lr$), and batch size ($b$) for all models. For the implementation, we use the Hugging Face transformers framework (Wolf et al., 2020).

[4]We post-process the results and automatically correct minor alignment errors, for example, false tag orders.

[5]https://github.com/dbmdz/berts

| model | train | test | precision | recall | $F_1$ |
|---|---|---|---|---|---|
| 0shot | NYT-0 | UMBL | 0.911 | 0.675 | 0.776 |
| | | ZEIT | 0.926 | 0.726 | 0.814 |
| | | COMB | 0.881 | 0.695 | 0.777 |
| | NYT-50 | UMBL | 0.890 | 0.306 | 0.456 |
| | | ZEIT | 0.867 | 0.429 | 0.573 |
| | | COMB | 0.869 | 0.354 | 0.503 |
| de2en | NYT-0 | UMBL | 0.809 | 0.599 | 0.689 |
| | | ZEIT | 0.822 | 0.642 | 0.721 |
| | | COMB | 0.780 | 0.616 | 0.688 |
| | NYT-50 | UMBL | 0.824 | 0.550 | 0.660 |
| | | ZEIT | 0.836 | 0.624 | 0.714 |
| | | COMB | 0.818 | 0.579 | 0.678 |
| en2de | NYT-0 | UMBL | 0.915 | 0.835 | 0.873 |
| | | ZEIT | 0.931 | 0.864 | 0.896 |
| | | COMB | 0.865 | 0.846 | 0.855 |
| | | ROB | 0.532 | 1.000 | 0.695 |
| | NYT-50 | UMBL | 0.907 | 0.803 | 0.852 |
| | | ZEIT | 0.936 | 0.843 | 0.887 |
| | | COMB | 0.887 | 0.818 | 0.851 |
| | | ROB | 0.574 | 0.794 | 0.667 |

Table 3: Performance on all test datasets using micro-average over all VA tags. "COMB" shows the scores for all test datasets (including NEG1, NEG2 and NEG3) combined. "ROB" shows the scores of the robustness study which is only conducted for the en2de model.

**0shot:** We fine-tune the XLM-RoBERTa base model that has 12 transformer blocks, 12 attention heads, a hidden size of 768 and ca. 270 million parameters ($e = 4, b = 16, lr = 5 \cdot 10^{-5}$).

**de2en:** We apply the German-to-English model from Ng et al. (2019) to translate the test datasets using FAIRSEQ. We align the words with SimAlign (Jalili Sabet et al., 2020), a word aligner which uses static and contextualized embeddings. We fine-tune the multilingual BERT base model (Devlin et al., 2018) (mBERT) ($e = 4, b = 64, lr = 3 \cdot 10^{-5}$).

**en2de:** We use the same tools as for de2en, but apply the English-to-German model from Ng et al. (2019) to translate the training data and apply SimAlign to project annotations. We fine-tune a German model (dbmdz/bert-base-german-cased) ($e = 3, b = 16, lr = 3 \cdot 10^{-5}$).

### 5.2 Experimental Evaluation

#### 5.2.1 Evaluation on annotated data

Table 3 shows the results. The en2de model outperforms the other two models by large margins on all test datasets. This comes as a surprise, we expected larger errors in the translation and the tag alignments of the training data. The results show only little differences compared with monolingual

approaches from Schwab et al. (2019) and Schwab et al. (2022). This is remarkable as the test datasets are much more diverse in terms of syntax and entity usage. For all models, it holds that they had better results on the ZEIT dataset compared to the UMBL dataset. One reason is the syntactic diversity of VA expressions UMBL contains. Most false negative errors were VA expressions that had different syntactic structure, like Example 2 in Table 1. In the negative samples, en2de predicted most false negative errors, but no more than 61 false tags in all 9,000 instances. Most tags were falsely predicted in sample NEG2. The addition of unlabeled data had a huge impact on the 0shot model where the performance dropped heavily. The other two models showed almost no difference.

#### 5.2.2 Robustness study

We conduct a study of our best performing model, en2de, trained with NYT0 and NYT50, respectively, on a sample of unlabeled real-world data. 1,000,000 random sentences of the TAZ corpus (introduced in Section 3) are tagged by the model. We analyze the predictions following Schwab et al. (2022): For each predicted tag, the tagger returns a prediction score. The tag with the highest score is the prediction. We compute the difference of the highest and second highest score which we interpret as a confidence score for the respective prediction. For all words of a sentence the confidence scores are averaged to represent the confidence of the prediction for a sentence. As in (Schwab et al., 2022), we take the 30 most confident predictions including at least one source and one modifier prediction tag (positive, i.e., a predicted VA expression) and 30 without those predictions (negative), as well as the 30 most unconfident (15 pos, 15 neg) to get the same share in the evaluation which we annotated manually. Table 3 indicates that the model has more problems on tagging real-world texts, it looses around .16 (NYT-0) and .18 (NYT-50) in $F_1$ compared to the COMB dataset. Still, the results are reasonable referring to the complexity of the task.

#### 5.2.3 Error Analysis

Analyzing the false positive prediction errors of our best model, en2de, it stands out that the model overfits to sentences that show syntactic patterns similar to the syntax of VA expressions in the training data. For example, in the sentence "Mike Stern ist das Raubein unter den Jazzgitaristen."[6] the model falsely tagged 'Mike Stern' as target, 'Raubein' as source and 'Jazzgitaristen' as modifier. A similar example is "AIDS – Super-Gau der Gentechnologie?"[7] where 'Super-Gau' was tagged as source and 'Gentechnologie' as modifier. In both examples, syntax and even semantic dependencies are close to the definition of VA. In particular, a common noun like 'Raubein' (or 'Super-Gau') is used in a specific environment, 'Jazzgitaristen' ('Gentechnologie', respectively). But as those nouns are no named entities and only the literal meaning of the words is used, the phrases cannot be VA expressions.

On the other hand, the false negative errors appeared mainly when the syntax of the VA expressions was new to the model, that is, it did not appear in the training data. Specifically, the modifier of VA expressions in the test data appeared before the source, for example, in "Wir brauchen einen neuen Don Quijote.", "Der russische James Bond heißt Stierlitz.", or "Eine griechische Cathy Freeman zu haben wäre nicht schlecht."[8]. In these cases, the model did not tag any word as part of a VA chunk but it should have tagged 'new', 'Greek' and 'Russian' as modifiers, 'Don Quijote', 'James Bond' and 'Cathy Freeman' as source, and 'Stierlitz' as target in the second sentence, whereas the other sentences do not have a target. This is a limitation of the model, even though in some other of these specific expressions, it tagged the chunks correctly.

## 6 Discussion

We analyzed cross-lingual VA extraction using English as source language and German as target language with limited annotated data. Our models achieve strong results which are even comparable to monolingual approaches like Schwab et al. (2022), also in the robustness study. Translating the training data to the target language worked best.

One goal for the future is to make more use of the semantics, even though this is a whole new problem as the examples have to be analyzed much deeper. Also, the generation of VA is a task we want to follow.

[6] "Mike Stern is the roughneck of jazz guitarists."

[7] "AIDS – Worst-case scenario of Genetic Engineering?"

[8] "We need a new Don Quijote.", "The Russian James Bond is called Stierlitz.", "Having a Greek Cathy Freeman would not be bad."

## References


1990. TRAVEL ADVISORY. *The New York Times*.

2005. Taz-archiv.

2013. Unter Wetterfröschen. *Der Spiegel*.

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. *arXiv preprint arXiv:2203.14139*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

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.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Sam Roberts. 1992. METRO MATTERS; Spirit of Newburgh Past Haunts Political Present. *The New York Times*.

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

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-IJCNLP)*, pages 6238–6243, Hong Kong, China. Association for Computational Linguistics.

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*. (in print).

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 (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.

Martin Víta. 2020. Cross-lingual metaphor paraphrase detection – experimental corpus and baselines. In *Information and Software Technologies*, pages 345–356, Cham. Springer International Publishing.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.