Triple Detection in German Verb-based Sentiment Inference: The Case of Novel Verbs

Dylan Massey
Department of Computational Linguistics, University of Zurich
Andreasstrasse 15, 8050 Zürich
dylan.massey@uzh.ch

Abstract

This short paper describes the evaluation of two neural models for their ability to detect sources, targets and verbal predicates as a step for enabling the full sentiment inference task, that is: Identifying whether a polar relation (*in-favour-of*, *against*) holds between two entities in a given sentence and which verb mediated the relation. The models are trained and evaluated on a silver standard generated by a rule-based system for sentiment inference on German text. We are mainly interested in the research question whether neural models are able to generalize to novel, previously unseen verb constellations and, thus, might make rule-based approaches superfluous. We find that one examined architecture, a simple transformers-based approach, achieves an F1-score of 85.2% on full triple detection.

1 Introduction

The central task in sentiment inference is to identify the proponents and opponents given a particular text. Polar verbs play a crucial role here, they identify the polar relation that holds between a source and target, which are both realized through particular semantic roles of the verb. In *The police man killed the aggressor with his weapon* there is an *against* relation: the police man as the source acts against the aggressor being the target. The presence of such a relation might be represented as a label (against) over the triple (police, kill, aggressor). However, not every instantiation of a verb should be interpreted that way. Take *The knife killed the aggressor*, an instrument-subject verb alternation. Here *knife* is not the source of an against relation, because it is not an actor. Adequate extraction of polar relations from sentences requires the consideration of selectional restrictions of verb frames identifying *knife* as inanimate.

In this paper we investigate whether a customized neural model, directly trained on a silver standard of triples is able to solve this task: finding the source, verb, and target. An exemplary triple is (X, loves, Y), which *might* indicate a favourable attitude of X towards Y. The source is denoted by X and the target by Y. The source is the origin of the (potentially inferred) sentiment described and the target the one towards whom the sentiment is directed. We are especially interested in the question whether the two investigated neural approaches and their underlying models generalise to verbs not seen in the training phase. Since for German no gold standard is available a silver standard was generated by a rule-based system (Klenner et al., 2017) whose precision is reasonable, but which is expected to have low recall due to lexical gaps.

2 Related Work

Sentiment Analysis is concerned with the elicitation of affective signal in text. While text types such as product reviews contain highly opinionated expressions, are short and profit from assignment of a polarity value in a range $[-1, 1]$, other text types such as newspaper articles, can benefit from more granular analysis. Approaches concerned with more granular, intra-textual elicitation of sentiments have been termed fine-grained sentiment analysis, aspect-based sentiment analysis, or more recently sentiment inference and structured sentiment analysis[1].

With the performance of approaches subsuming transformer-based contextual word embeddings (Vaswani et al., 2017; Devlin et al., 2019) tasks related to the elicitation of fine-grained sentiment in text have profited too. A commonly found distinction made by approaches is the separation between entity recognition (ER) and subsequent classification of which relation holds between them. The former has given rise to the subtask of Opinion Role Labelling (ORL). ORL has the goal to

[1]For structured sentiment analysis cf. Barnes et al. (2021)

identify source and target given a polar sentence (Bamberg et al., 2022, p. 112). Previous work has also considered the more extensive task of additionally identifying a cue expression inducing sentiments between or towards entities (Yang and Cardie, 2013; Choi and Wiebe, 2014; Katiyar and Cardie, 2016). In the ORL-only setting and more recently in German, Bamberg et al. (2022) achieve state-of-the-art performance on the IGGSA-Steps datasets (Ruppenhofer and Struss, 2016), they too, use a transformers-based approach. Deng and Wiebe (2015) are the first to present a fine-grained corpus of texts for sentiment analysis. Reschke and Anand (2011) introduce the idea of verbs for an implicit sentiment analysis in English.

3 Rule-based Silver Standard

The rule-based system described in Klenner et al. (2017) uses a verb lexicon[2] for sentiment inference. For the instantiation of a verb, various restrictions must be satisfied. Table 1 shows one polar frame for the German verb *sorgen für* (*care for*).

1 dependency label	subj	pp-obj
2 lexical restriction	-	prep=für
3 selectional restriction	+animate	+animate
4 polar role	source	target
5 polar relation	in favour	-

Table 1: Frame of *sorgen für* (Eng. care for)

For this reading the restrictions 1, 2 and 3 must hold: particular dependency labels, lexical restrictions (preposition für, Eng. *for*) and animacy requirements must be met. Then the polar assignments (4 and 5) can be made, namely that the subject is the source and the object the target of a in favour relation. A dependency parser (Sennrich et al., 2009) and an animacy classifier (Klenner and Göhring, 2022) are used to enforce those restrictions. Due to the restrictive verb instantiation policy, precision of the system is reasonable. We manually evaluated the output of 210 sentences and found a precision of 79.0%, a recall of 78.9%. However lexical gaps (missing verbs) and not modelled polar versions of a verb are expected to affect recall. Here neural models come into play, which might be able to generalise to unseen verbs. However, these models not only should learn applicable verbal profiles, but also the restrictions that guide the instantiations (e.g. selectional restrictions) of source and target. For this, a large gold standard is needed. Since no such gold standard for German is available, we propose to create a silver standard on the basis of the output of Klenner et al. (2017), which - as we have argued - has proved to have a reasonable performance. We used the demo system[3] of Klenner et al. (2017) to produce a silver standard.

The data which is used to generate the silver standard dataset is from the Swiss Media Database (SMD)[4]. In total 266,647 news articles from major Swiss news outlets within the date range from January 2018 until November 2022 were downloaded, cleaned and passed through the rule-based sentiment inference system. We omit sentences which do not include both, a source and a target and where either or both are pronouns. Both settings would require means of resolution over text surpassing the sentence-level, which is outside the scope of our work. Along with the aforementioned source, target, verb triples, the rule-based system also generates a label that signifies a positive or negative relation between the source and target depending on the constellation between the three entities. We also include an equal amount of "neutral" triples all of which contain verbs however that can potentially be charged and are thus part of the verb lexicon. Including triples that contain polar verbs in neutral constellations can help reduce detection of false positives and therefore lead to a more adequate performance in sentiment inference and analysis systems (Wilson, 2008, p. 181). Although these triples are not per se triples that induce a polar relation and we cannot talk of a source and target in such instances, we keep the terms for simplicity.

4 Neural Models for Polar Triple Detection

We have so far discussed the task of sentiment inference and the importance of identifying the verbal predicate and the roles the predicate casts on its participants as a first step. Contrary to lexicon-based systems, neural approaches handle out-of-vocabulary words at inference and might cope well with unseen verb constellations. For example, if *X loves Y* is within the vocabulary of the lexicon,

[2]The lexicon is available from https://www.cl.uzh.ch/en/texttechnologies/research/opinionmining/sentiment-inference.html

[3]The demo is available under https://pub.cl.uzh.ch/demo/stancer/index.py

[4]https://smd.ch/en/home

then we might infer that X is positive towards Y, given X := subj and Y := obj. Similarly, *X adores Y* might not trigger if *adore* is not in the lexicon.

Our interest lies on the polar relations that are verb-mediated and span **between** two proper textual entities. The goal of our neural task is to (a) detect entities (verb and its fillers) on the sentence-level and (b) how the found non-verbal entities relate to each other. More precisely, our goal is to detect all triples ρ from some sentence x where ρ consists of a source s, a verbal predicate v and a target t. In the present paper **we shall only be concerned with step (a), that is the detection of the verbal predicate and its fillers**, since initial experiments revealed that the downstream performance of (b) is greatly affected by the ability to extract salient triples from the sentence (especially the verb).

5 Method

We consider two systems for sentiment inference. System 1 (S1) is devised by Zhong and Chen (2021) who follow a two-step approach for the task of relation extraction (RE), which we will repurpose for sentiment inference and abbreviate as ERRE[5]. They train entity model and relation classifier independently of each other. The entity-marked sentences serve as input to the relation model. We base our implementation closely on the tutorial provided by Pal (2022), who references Zhong and Chen (2021), but modify the approach to handle ternary relations aligning our task. Both approaches use `XLM-R` by Conneau et al. (2020) as pre-trained base. For the relation model of the ERRE we use the pre-trained `bert-base-german-cased`, which has shown performant in German language settings. System 2 (S2) is proposed by Samuel et al. (2022) for structured sentiment analysis, where the goal is to extract and polarly relate subjectivity cues to sources, targets. We abbreviate System 2 as PERIN.

Initial experiments revealed that final task performance is affected most dramatically by ER performance relative to downstream RE. Thus, the focus of the present paper is on ER performance. Since ERRE relies on two independent models, ER performance can be easily evaluated. Its entity model is based on `AutoModelForTokenClassification` (Wolf et al., 2020). For PERIN it is not possible to fully decouple entity recognition capabilities from the full task since training occurs end-to-end. Therefore we train and on the full dataset (including on in-favour and against relations) but evaluate only ER performance.

To test the generalization capabilities of our proposed neural models we split test, validation and training data twice: once randomly (RAND) and once based on the restriction that all triples that have some verb v can only either all be in the training dataset or in the test dataset (MEVG[6]. The MEVG setting simulates the worst-case scenario where all verbs are novel at inference and reveals models generalisation capabilities to unseen contexts.

RAND contains 28,072 train, 6,004 validation and 6,082 test sentences. Label frequencies were balanced using a disproportionately stratified sampling meaning that we have equal representation of every class - in favour, against, neutral - in our dataset. RAND consists of 460 unique verbs in total. MEVG contains 30,032 training and 5,063 validation and testing sentences. The training dataset contains 333 unique verbs and validation and testing (where we allow overlap) contain 138 unique verbs.

As evaluation metrics, we rely on Barnes et al. (2022), who evaluate the performance on the F1-score, precision and accuracy for each respective element of the n-tuple as well full tuple precision and accuracy. The precision and recall are provided in the appendix A.2. Hyperparameters are addressed in appendix A.1.

6 Results

We evaluate both approaches on their ability to correctly identify the verb and potential sources and targets and conclude that the entity model of the ERRE approach performs best on our generated silver standard.

Table 2 illustrates the F1-scores for the individual entity extraction scores and the performance for the combined triple extraction, F1_ρ, on the test dataset. The performance drops from RAND to the MEVG setting where only verbs not part of the training dataset are in the test dataset. The drop is higher for PERIN (S_2) than for ERRE (S_1) and it is more drastic for ERRE in verb (F1_V) (15.1%) and target detection (F1_t) (15.3%) than it is for source detection (F1_S) (7.9%). The decrease in (whole) triple classification performance (F1_ρ) is

[5]Short for Entity Recognition and Relation Extraction

[6]Mutually exclusive verbs groups splitting

System	Split	$F1_S$	$F1_t$	$F1_V$	$F1_\rho$
S_1	RAND	**90.7**	**88.7**	**96.8**	**85.2**
S_2	RAND	89.5	85.8	95.9	83.9
S_1	MEVG	82.8	73.4	81.7	65.0
S_2	MEVG	47.3	42.1	47.0	41.8

Table 2: Triple recognition capabilities for source ($F1_S$), target ($F1_t$) and verb ($F1_V$) of S_1 (ERRE) and S_2 (PERIN) on the test dataset in % depending on whether verbs in the test dataset were randomly overlap (RAND) or were mutually exclusive to the training dataset (MEVG).

19.8% for ERRE (from 85.2% to 65%) and 42.1% (from 83.9% to 41.8%) for PERIN. PERIN performs worse than ERRE under unseen constellations. Unexpectedly performance is not only low on verb detection, but similar also on source and target detection performance. ERRE, on the other hand, not only has still (under MEVG) relatively better performance in verb detection (81.7%), but also a triple F1 score of 65% . The loss in triple classification from entity recognition appears rather attributable to the 12.4% drop in target classification (from 85.8% to 73.4%) rather than verb identification.

In order to get a better understanding of the quality of the silver standard and the reproductive power of the neural models, a randomly sampled set of 210 sentences from the silver standard were manually annotated by a single annotator using the Universal Data Tool by Ibarluzea (2021). This still can not be regarded as an analytic gold standard, since these 210 sentences were selected on the basis that at least one verb of the rule-based system's lexicon was present (100% sentence recall). However, since the objective of the current paper is to evaluate the verb generalisation power of neural models, we argue that this is a reasonable initial setting (we could call it a verb-biased gold standard **gold*) for the sake of investigation. Since the verb was pre-supposed (and shown to the annotator), we excluded it from the evaluation. We only carried out opinion source identification, i.e. how accurate the silver standard is in terms of identification of potential sources and targets. Results are visible in Table 3.

Knowing the limitations (we only have **gold*) we nonetheless can say that the silver standard appears to satisfy performance wrt. opinion role labelling and acts as viable resource for training neural models. The manual annotation of S_0 revealed a $F1_S$ of 87.3%, a $F1_t$ of 85.9% and a $F1_{s,t}$ of 78.9% ($F1_{s,t}$ indicates pairs of source and target). Both neural approaches reproduced these results: Trained on the silver standard, they reach the same performance wrt. to **gold* (a real gold standard, though verb-biased) as S_0. For more precise results including the detailed precision and recall for all the individual and the combined components we refer to the appendix in section A.2. Nevertheless, the results are - due to the restriction that only the opinion roles of the given verbs are considered - too high and cannot be compared to full opinion role detection as e.g. done in Bamberg et al. (2022). There the results are 10 to 15% lower.

System	$F1_S$	$F1_t$	$F1_{s,t}$
S_0	87.3	85.9	78.9
S_1	88.3	84.3	77.2
S_2	88.7	84.5	77.1

Table 3: Comparison of the silver standard (S_0), the ERRE (S_1) and PERIN (S_2) models with manually annotated sentences (**gold*). All numbers are in %.

7 Conclusion

The research hypothesis of this short paper was that neural models are able to deal with novel and unseen verbs not encountered during training in the context of sentiment inference. This is crucial where the polar verb directly mediates a polar relation (in favour, against) if used in an affirmative, factual sentence. On the basis of a rule-based system, a silver standard was generated for the training and evaluation of two neural models. The empirical settings comprised a worst-case scenario where the verbs of the training and test set are mutually exclusive. In this setting, the performance of one of the systems, though decreased, still was reasonably good showing that generalization at the verb-level has taken place. This learned neural model is as good as the rule-based system on a small gold evaluation, but can also deal with novel cases which the rule-based system under no conditions could achieve. The rule-based system, thus, is superfluous. With increasingly powerful large language models we devise as future work to inves-

tigate prompting techniques as data augmentation strategy for our current models, as well as a direct approach using large language models.

References

Laura Bamberg, Ines Rehbein, and Simone Ponzetto. 2022. Improved Opinion Role Labelling in Parliamentary Debates. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 110–120, Potsdam, Germany. KONVENS 2022 Organizers.

Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2021. Structured Sentiment Analysis as Dependency Graph Parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3387–3402, Online. Association for Computational Linguistics.

Jeremy Barnes, Laura Oberlaender, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. SemEval 2022 Task 10: Structured Sentiment Analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1280–1295, Seattle, United States. Association for Computational Linguistics.

Yoonjung Choi and Janyce Wiebe. 2014. +/-effectwordnet: Sense-level lexicon acquisition for opinion inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, SIGDAT*, pages 1181–1191.

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, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Lingjia Deng and Janyce Wiebe. 2015. MPQA 3.0: An entity/event-level sentiment corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1323–1328, Denver, Colorado. Association for Computational Linguistics.

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.

Severin Ibarluzea. 2021. Universal Data Tool.

Arzoo Katiyar and Claire Cardie. 2016. Investigating LSTMs for Joint Extraction of Opinion Entities and Relations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 919–929, Berlin, Germany. Association for Computational Linguistics.

Manfred Klenner and Anne Göhring. 2022. Animacy denoting german nouns: Annotation and classification. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1360–1364, Marseille, France. European Language Resources Association (ELRA).

Manfred Klenner, Don Tuggener, and Simon Clematide. 2017. Stance detection in Facebook posts of a German right-wing party. In *LSDSem 2017/LSD-Sem Linking Models of Lexical, Sentential and Discourse-level Semantics*.

Sujit Pal. 2022. Transformer Based Approaches to Named Entity Recognition (NER) and Relationship Extraction (RE).

Kevin Reschke and Pranav Anand. 2011. Extracting contextual evaluativity. In *Proceedings of the Ninth International Conference on Computational Semantics*, pages 370–374.

Josef Ruppenhofer and Julia Maria Struss. 2016. IGGSA-STEPS: Shared Task on Source and Target Extraction from Political Speeches. 29(1):33 – 46. Place: Regensburg Publisher: Gesellschaft für Sprachtechnologie und Computerlinguistik.

David Samuel, Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2022. Direct parsing to sentiment graphs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 470–478. Association for Computational Linguistics.

Rico Sennrich, Gerold Schneider, Martin Volk, and Martin Warin. 2009. A new hybrid dependency parser for German. In *Proceedings of the German Society for Computational Linguistics and Language Technology*, pages 115–124.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Theresa Ann Wilson. 2008. Fine-grained Subjectivity and Sentiment Analysis: Recognizing the intensity, polarity, and attitudes of private states. Doctoral Dissertation, University of Pittsburgh. (Unpublished).

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 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.

Bishan Yang and Claire Cardie. 2013. Joint Inference for Fine-grained Opinion Extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1640–1649, Sofia, Bulgaria. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 50–61. Association for Computational Linguistics.

A Appendix

A.1 Hyperparameters

For the entity model part of ERRE we rely on sensible defaults. This amounts to a batch size of 16 for both training and development/test sets, 3 epochs, a learning rate of 2×10^{-5} and a weight decay of .01. Cross-entropy is used as a loss function and Adam serves as the optimizer.

For PERIN we use AdamW as the optimizer, a linear scheduler (without warmup). A weight decay of 1×10^{-2} is set, while the learning rate is set to 2×10^{-2}. The model is trained again for 3 epochs. Cross-entropy is the loss function.

A.2 Performance

Full performances for the silver standard test dataset in comparison to predictions of PERIN:

```
Source Precision: 0.932
Source Recall: 0.861
Source F1: 0.895
Target Precision: 0.901
Target Recall: 0.819
Target F1: 0.858
Verb Precision: 0.959
Verb Recall: 0.959
Verb F1: 0.959
Tuple Precision: 0.836
Tuple Recall: 0.842
Tuple F1: 0.839
```

Full performances for the manual annotations in comparison to the predictions of ERRE:

```
Source Precision: 0.895
Source Recall: 0.920
Source F1: 0.907
Target Precision: 0.903
Target Recall: 0.871
Target F1: 0.887
Verb Precision: 0.969
Verb Recall: 0.966
Verb F1: 0.968
Tuple Precision: 0.853
Tuple Recall: 0.851
Tuple F1: 0.852
```

Full performances for the manual annotations in comparison to the silver standard:

```
Source Precision: 0.896
Source Recall: 0.851
Source F1: 0.873
Target Precision: 0.857
Target Recall: 0.861
Target F1: 0.859
Tuple Precision: 0.790
Tuple Recall: 0.789
Tuple F1: 0.789
```

Full performances for the manual annotations in comparison to the entity model of the ERRE system:

```
Source Precision: 0.877
Source Recall: 0.889
Source F1: 0.883
Target Precision: 0.833
Target Recall: 0.854
Target F1: 0.843
Tuple Precision: 0.771
Tuple Recall: 0.774
Tuple F1: 0.772
```

Full performances for the manual annotations in comparison to the entity model of the PERIN system:

```
Source Precision: 0.877
Source Recall: 0.898
Source F1: 0.887
Target Precision: 0.837
Target Recall: 0.854
Target F1: 0.845
Tuple Precision: 0.771
Tuple Recall: 0.771
Tuple F1: 0.771
```