Getting Hold of Villains and other Rogues

Klenner, Manfred ; Göhring, Anne ; Conrad, Sophia

Getting Hold of Villains and other Rogues

Manfred Klenner & Anne Göhring & Sophia Conrad
Department of Computational Linguistics
{klenner|goehring|conrad}@cl.uzh.ch

Abstract

In this paper, we introduce the first corpus specifying negative entities within sentences. We discuss indicators for their presence, namely particular verbs, but also the linguistic conditions when their prediction should be suppressed. We further show that a fine-tuned BERT-based baseline model outperforms an over-generating rule-based approach which is not aware of these further restrictions. If a perfect filter were applied, both would be on par.

1 Introduction

In online media including social media, the world is often conceptualized as being divided into beneficiaries and benefactors, victims and villains. For quite some time, the most interesting questions seem to have been: Who is to blame and who benefits most. In this work, we strive to create a dataset and a first model to answer the first of these questions, i.e. to identify the villains in texts. But what is a villain, anyway? Are we compelled to reveal our moral convictions in order to answer this question? A murderer, a cheater, a liar seem to be clear cases. But what about white lies and the cheating in a card game? We could introduce a severeness score in order to quantify the villainousness grade.

In this paper, we describe our annotation efforts to create a corpus of sentences that comprises at least one entity that realises a negative (semantic) role. The filler of a negative semantic role might be a person, organization etc. But it also might be an event or even a non-animate physical object. Especially in metonymic constructions, non-animate fillers are to be expected. Although we also have started to annotate the strength of negativity, in this short paper we focus on the language usage that gives rise to the assignment of negative roles per se. In the second part of the paper we discuss two models: a rule-based and a BERT-based one.

2 Phenomena to be considered

The goal of our annotation is to identify those entities of a sentence that occupy a negative semantic role. A number of constructions can be used to take a negative perspective on some entity. One can do it explicitly by a noun phrase (*the lies of the president*), a predicative construction (*he is a liar*) or by using a verb who implies a negative actor (*He vilifies the people*). In this paper, we focus on verbs. It turns out, though, that not every usage of such a verb assigns a negative role. Only if the situation at hand is factual, then a negative role actually indicates a villain (Klenner and Clematide, 2016). Also ambiguity has to be taken into account. We, thus, are talking about a probability distribution, depending on various grammatical parameters. Before we have a more detailed look at this grammatical means, please note that quite a couple of verbs do have negative roles especially at the actor position. We have identified about 400 for the German language (Klenner and Amsler, 2016). Among them are verbs that indicate a crime (e.g. to murder, to kill, to injure, to torture ...), but also verbs like to vilify, to rebuff, to lie, to cheat, to mock, to demoralize, to prejudge and so on. Most of the time, the subject of the verbs bears the negative role.

As we said, metonymic reference has to be taken into account. Besides classical cases of metonymy like *producer for product* (e.g. *Pynchon is hard to read*), we also consider all references to be metonymic when humans are involved, e.g. *This agreement destroys our hope*.

There are a number of grammatical means (see Fig. 1) that indicate non-factuality and thus block the assignment of negative roles.

In reported speech (1), the actor of the reported event (*China*) is blocked. Subjunctive mood (2) inhibits the inference (*agreement*) since nothing has happened, which is also true for future tense (3). For verbs that have a theme dependent nega-

1. reported speech: *He said China was responsible for the virus*
2. subjunctive mood: *This agreement would destroy our hope*
3. future tense: *He will deny his guilt*
4. pronoun underspecification: *They admit it*
5. modal verbs: *The UN must invade*
6. modal adverbs: *He possibly is lying*
7. negation: *He has never cheated the people*
8. conditional constructions: *If he lies, the people won't elect him*
9. reflexive usages: *He cheats himself*
10. different reading: *He hurts the deadline*

Figure 1: Inference Blocker

tive actor assignment (like *to admit a mere/serious mistake/crime*), an unresolved pronoun (4) blocks inference. Some modal verbs (5) prevent the assignment of negativity (*UN*) as do adverbs (6) like *possibly*. Negation (7) also acts as a plug for such inferences, as well conditional statements (8). We also argue that the reflexive use of these verbs is not indicating a negative actor (9). Harder to detect are cases where the right reading should suppress the assignment of a negative actor (10).

In traditional machine learning we would use the items from Fig. 1 as features. A rule-based system could try to use them as filters. In a Deep Learning scenario, e.g. a BERT-based model, we could hope that the fine-tuning process will be sufficient to learn the regularities.

3 Annotated Corpus

As source for sentences that might have a negative role, we selected 1300 sentences from two corpora[1]. The first one is a German newspaper corpus called TuebaDZ (Telljohann et al., 2009) comprising more than 100,000 sentences (publically available) and the second one are Facebook posts of a German right-wing party (AfD) with more than 300,000 sentences[2]. The AfD texts also contain offensive language. The TuebaDZ data, on the other hand, comes from a left-oriented newspaper. We deliberately have chosen two different world views in order to have a broader range of examples.

We generated the candidate sentences by the following procedure: we parsed the sentences with the German ParZu parser (Sennrich et al., 2009) and then for those sentences that had a verb from our lexicon, we extracted the predicate argument structures (as a preprocessing step of our rule-based system, see (Klenner et al., 2017)) from the dependency parse trees. Finally, we identified the agent position (ARG0) and suggested it to be a negative actor. Given

Unser Land wird von den Medien zerstört

which translates to *Our country is being destroyed by the media*, the predicate argument structure (as a formula in Predicate Logic) is *destroy(media,country)*. From this, *media* was extracted to denote a negative actor *negative_role_filler(media)*. The two annotators were presented with the full sentence and had to determine whether the suggested negative actor actually is one. Moreover, the strength of negativity had to be determined on the basis of a scale from 1 (low) up to 3 (high). A zero means false positive.

In the course of the annotation, we removed a couple of sentences, because no actor was found by the predication extractor. We ended up with 1260 sentences. 460 cases are false positives, i.e. the found actor was not a negative actor, exactly 800 were true positives. We had a closer look at the reasons for the false positives, i.e. how the criteria from Fig. 1 are distributed. Only 4 cases would need coreference resolution, 18 are errors based on negation, 19 cases were future tense, 19 reported speech, 38 subjunctive mood, 48 reflexive usage, 59 conditional forms and 162 were wrong readings. We also had a number of parsing errors, namely 93 (wrong candidate). From 460 cases of false positives, thus, 183 cases (39.78 %) can be blocked by a perfect filter, coreference, negation, parsing errors and cases of wrong readings due to ambiguity are out of reach.

Our inter-annotator agreement is a Kappa score of 0.78: i.e. whether the annotators agreed that a noun candidate was really a negative actor or a false positive.

[1]The annotated data is available, just contact the first author.

[2]The data is publically avaiable on request - please contact the first author.

4 Experiments

4.1 Rule-based Baseline

We used our rule-based system for sentiment inference[3] as a baseline. The system is verb-based and is designed to generate all pro (in favor of) and con (against) relations among the entities mentioned in a text. Moreover it indicates which discourse referents are negative actors and which receive a negative effect (see (Klenner et al., 2017) for the details). We just took the negative actors from the output and tested against our gold-standard. The system is over-generating: a rule triggers if a verb from the lexicon is found and the syntactic frame of the verb is met by the parse tree. We have not realised a filter (from Fig. 1) to block non-factual sentences from producing negative actors, but we give the hypothetical improvement the rule-based system would achieve if it was available (RB* in Tab. 1).

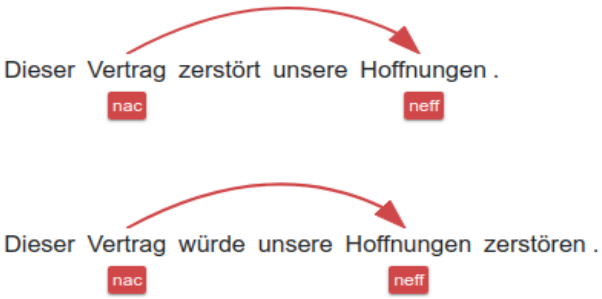

Figure 2: Top is right, bottom wrong

Figure 2 shows the result for the sentence *The contract destroys our hope* (top) and *The contract would destroy our hope* (subjunctive, bottom). The negative actor (*nac*) *Vertrag* (*contract*) stands in a con relation (red arc) with *Hoffnungen* (*hopes*) which receives a negative effect (*neff*). Only the top analysis is right, since the second sentence is not factual.

The advantage of such a rule-based system is transparency. The logic behind the predictions can be analysed, further refined, and applied to new verbs, if needed. The backside is that it remains brittle: lexical gaps (reduced recall) and erroneous parse trees (reduced precision) affect the performance. Moreover, if we are interested only in a well-performing system (in some end-to-end architecture), we do not necessarily need transparency. It might turn out that a neural approach is on par with or even outperforms it. Also, the next step in our research strives to automatically quantify the severeness of negativity of an actor. Here a regression analysis is a natural approach (given that enough training material is available which is current work).

[3] https://pub.cl.uzh.ch/demo/stancer/

4.2 BERT-based Model

We realised a straightforward neural model by fine-tuning a German BERT model[4]. Several runs with different test sets showed that the results only slightly vary.

We tried two scenarios. In the first one, the whole sentences together with the labels were given to the training procedure. The binary classification task then was to label the sentence either as 1 (true positive) or 0 (false positive). If 1, then we know that the candidate noun (ARG0, which is mostly the subject of the verb from our verb lexicon) is a negative actor, given 0, it is not. The results were not very promising. We achieved 61% precision and 52% recall. We, thus, stopped further experiments with this setting.

In the second and simpler setting, the training procedure just gets all words between ARG0 and its verb (including ARG0 and the verb). Due to German word order, it might be the other way round as well (every word between the verb and ARG0). Sometimes, a potential indicator word (e.g. an adverb) gets lost in these cases, but they are rare. To give an example of such a fragment:

"*er die Frauen immer wieder anschrie*" (*he yelled at the women again and again*)

ARG0 is *er (he)* and the verb from the lexicon is *anschreien (to yell)*. In Table 1 we provide the results of four runs with BERT (DL 1-4) and the single result from the application of the rule-based model (RB).

First of all, the RB model does not trigger on every sentence. The reason might be a missing verb subcategorization frame or a wrong dependency tree (the model only triggers if the verb frame from the lexicon is found). This explains the recall below 100%. The precision is lower than that of the DL model since the RB model is not able to identify class 0. Some of the predicted 1, thus, are 0. The recall of RB is higher than those of DL probably since no inference blocking mechanisms are imple-

[4] We use the Transformer library from HuggingFace (Wolf et al., 2020) and the BERT model made publicly available at https://huggingface.co/bert-base-german-cased

	precision	recall	f-measure
RB	64.87	88.04	74.83
RB*	71.83	88.04	78.59
DL 1	72.77	84.75	78.31
DL 2	74.86	79.87	77.28
DL 3	72.02	84.75	77.87
DL 4	72.82	86.58	79.10
DL mean	73.12	83.99	78.14
DL std	1.05	2.49	0.66

Table 1: Rule-based (RB) versus BERT-based (DL): label 1

	precision	recall	f-measure
DL 1	42.30	60.12	49.65
DL 2	51.11	58.22	54.43
DL 3	40.00	59.01	47.68
DL 4	41.11	62.71	49.66
DL mean	43.63	60.02	50.36
DL std	4.39	1.70	2.49

Table 2: Results of BERT-based (DL): label 0

mented and, thus, more is predicted. RB* gives the results if a perfect filter was applied. Out of the 183 filtered out cases[5], 102 have triggered an inference. If we reduce the number of found cases (the denominator of precision) by 102, precision goes up to 71.83% and the f-measure raises to 78.59%. Both approaches were on par, then.

The DL model has - to a certain extend - learned that some examples belong to the category 0. Table 2 shows the DL results for the label 0. They are worse than those for 1. This might stem from the truncation which sometimes cuts away too much.

5 Related Work

Our task, detecting negative actors, is somehow related to the task of opinion role identification (see e.g. (Wiegand et al., 2019)), where the goal is to identify the source (our case) and the target of an opinion event expressed by a sentence. However, our task is more specific and more general at the same time. We are interested in opinion sources that also are conceptualized as negative actors as in *He vilifies the people*. But we not only are looking at opinion sources but are also interested in any event source (or actor) that is negatively connotated through the sentence (e.g. *He deliberately injured others*). There is a superficial similarity with the work of (Wiegand et al., 2016) where also a rule-based approach was used. Our rule-based system is described in (Klenner et al., 2017). Among others, it also produces predictions of negative actors. The system uses a large verb lexicon where each verb is specified according to its various syntactic frames and where the frame elements are further specified with respect to their polar roles (e.g. negative actor) and the pro or con relation among each other. The shortcoming of the system clearly is that it is over-generating, it only partially is able to identify non-factuality and has no means to distinguish relevant from irrelevant readings of a verb.

We have also tried to find related work in the fields of stance detection and even argumentation mining. But we are not aware of any approaches that directly focuses as we do on that task, neither for German nor for any other language. The field of emotion classification is relevant, as negative actors might evoke strong emotions. Inspired by (Oberländer et al., 2020)'s paper title, we want to investigate "which semantic roles enable machine learning to infer" the negative sentiments towards agent entities. Based on cognitive appraisal theories, the corpus of thousand sentences described in (Hofmann et al., 2020) explicit the link between emotions caused by events and the appraisal dimensions. Our ongoing attempt to quantify the severeness of negativity involved might benefit from a closer look at the emotional side. (Bostan et al., 2020) annotated emotions in English news headlines via crowdsourcing, together with semantic roles and the reader's perception. To gain more insights into the severeness dimension, crowdsourcing could be a good way in order to come to a more representative since larger corpus.

6 Conclusion

We have introduced a dataset of 1260 actor-verb pairs (including their sentences) where each pair either identifies a negative actor of a factual situation described by the verb or the actor is not a negative actor mostly because the verb denotation - given the sentence - is non-factual. Factuality could in principle be determined on the basis of certain grammatical indicators, but other inference blockers are harder to identify (verb ambiguity). If the rule-based system had a basic (and perfect) filter, both approaches, DL and RB, are on par. We have shown that the neural models (DL) are able to learn the needed distinctions without relying on manual

[5] As discussed, 39.78 % of the 460 false positives could be detected by a simple filter.

feature engineering or manual filter specifications. The identification of negative actors might be useful for a system that detects offensive language or hate speech, where targets (e.g. migrants) quite often are being conceptualized as villains.

Future work will focus on the determination of the strength of negativity. Manually quantifying negativity of actors is an error prone task. Moreover, the actual strength value is not so crucial - it is the right ranking (is actor A more negative than actor B) that counts. We have started to experiment with a lexicon-based quantification metric (see (Clematide and Klenner, 2010) for the lexicon) that takes into account different types of sentiment specifications, e.g. appraisal categories (Martin and White, 2005) (judgement, emotion, apprehension words) and the classification of emotion words according to the base emotions they express (Plutchik, 1980).

7 Acknowledgements

Our work is supported by the Swiss National Foundation under the project number 105215_179302.

References

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France. European Language Resources Association.

Simon Clematide and Manfred Klenner. 2010. Evaluation and extension of a polarity lexicon for German. In *Proceedings of the First Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, pages 7–13.

Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. 2020. Appraisal theories for emotion classification in text. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 125–138, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Manfred Klenner and Michael Amsler. 2016. Sentiframes: A resource for verb-centered German sentiment inference. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Manfred Klenner and Simon Clematide. 2016. How factuality determines sentiment inferences. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 75–84, Berlin, Germany. Association for Computational Linguistics.

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*.

J. R. Martin and P. R. R. White. 2005. *Appraisal in English*. Palgrave, London.

Laura Ana Maria Oberländer, Kevin Reich, and Roman Klinger. 2020. Experiencers, stimuli, or targets: Which semantic roles enable machine learning to infer the emotions? In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, Barcelona, Spain (Online). Association for Computational Linguistics.

Robert Plutchik. 1980. *A general psychoevolutionary theory of emotion*. Academic press, NewYork.

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

Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. 2009. Stylebook for the Tübingen treebank of written German. Technical report, Universität Tübingen, Seminar für Sprachwissenschaft.

Michael Wiegand, Nadisha-Marie Aliman, Tatjana Anikina, Patrick Carroll, Margarita Chikobava, ErikHahn, Marina Haid, Katja Konig, Leonie Lapp, Artuur Leeuwenberg, Martin Wolf, and Maximilian Wolf. 2016. Saarland university's participation in the second shared task on source, subjective expression and target extraction from political speeches (steps-2016). In *Proceedings of IGGSA Shared Task Workshop, Bochumer Linguistische Arbeitsberichte*.

Michael Wiegand, Margarita Chikobava, and Josef Ruppenhofer. 2019. A supervised learning approach for the extraction of sources and targets from German text. In *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019, Erlangen, Germany, October 9-11, 2019*.

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.