Cracking a Walnut with a Sledgehammer: XLM-RoBERTa for German Verbal Idiom Disambiguation Tasks

Franziska Pannach, Tillmann Dönicke
Göttingen Centre for Digital Humanities
Georg-August Universität Göttingen
firstname.lastname@uni-goettingen.de

Abstract

This paper describes the efforts in solving the Shared Task on the Disambiguation of German Verbal Idioms at KONVENS 2021. It presents the team's efforts to extend the training data semi-automatically. The disambiguation task was solved using XLM-RoBERTa, which delivered the best results with 0.76 f1-Score on all tested non-idiomatic instances in the test set. The baseline model, a linear SVM, achieves 0.55 f1-Score. Furthermore, additional data was collected to enhance the training data set with respect to literal use of idiomatic expressions. While the baseline model improves slightly with additional training data, the XLM-RoBERTa model performs better when only the core training data is provided.

1 Introduction

Verbal idioms (VID) are a special type of multi word expression that serve as instruments of figurative language and contain a verbal construction, such as *to kill two birds with one stone*. Due to the verbal structure, these types of idioms embed naturally into sentences, while other types of metaphoric language may appear more likely in stand-alone sentences, such as *Time is money*. Therefore, identifying VIDs is a challenging task in natural language processing.

Some verbal idioms only occur in their figurative form, such as *to be born with a silver spoon in one's mouth*. However, in many cases, verbal idioms have a figurative and a literal meaning. Their disambiguation relies on the context they appear in.

In this paper, we present an approach to verbal idiom disambiguation for German idioms as a contribution to the Shared Task on the Disambiguation of German Verbal Idioms at KONVENS 2021. The additional data and jupyter notebook for the submission can be accessed online.[1]

The paper is structured as follows: Section 2 investigates the data and task definition. In Section 3, the algorithms for the baseline and final disambiguation is presented. Results can be found in Section 4. Conclusion and further work are discussed in Section 5.

2 Data and Task Description

The shared task data consists of approximately 9.9k samples in total which are divided into training, development and test set (see Table 1). Each sample consists of ID, type, label and text. The text usually consists of three sentences, where the verbal idiom appears in the middle sentence. Each token belonging to the verbal idiom is marked up by `<b>...</b>` tags. The aim of the shared task is to develop systems that take a text (including markup) as input and predict the corresponding label, i.e. LITERALLY and FIGURATIVELY, as output.

The proportion of the labels LITERALLY and FIGURATIVELY is about 1:5 in each of the three sets. A small proportion ($< 1\%$) of samples is further labelled with BOTH or UNDECIDABLE. However, these samples do not carry weight in the shared task, since the participating systems are evaluated on the performance on the LITERALLY class, i.e. how well a system can predict non-figurative examples of utterances that usually carry metaphorical meaning.

Although the training set contains about 6.9k samples, the number of types only amounts to 61. Examples for all of these types are also contained in the development and test set, and are considered *seen* in the evaluation. Both the development and the test set further contain examples for three

[1] https://gitlab.gwdg.de/franziska.pannach/shared-task-idioms-submission

	literally	figuratively	total	types
train	1,172	5,705	6,902	61
dev	264	1,214	1,488	64
test	265	1,238	1,511	64
extra	60	1,106	1,166	561

Table 1: The shared task's datasets and the additional dataset.

additional types that are considered *unseen* in the evaluation.

Since the number of types in general and instances of literal use of VID in particular are comparatively small (and it was not clear in the development phase of the shared task whether the test set will show the same overlap with the training set as the development set), we collected additional training data from various open sources.

First, we extracted all pages in the German **Wiktionary** that are tagged with "Redewendung" ('idiom'). The results were then filtered by two conditions:

1. The idiom must be of the form `(preposition? attribute* noun)+ adverb? verb+`, where {`?`, `*`, `+`} are used as in regular expressions. This step yields *verbal* idioms.

2. The idiom must not contain the word *wie* 'like'. This excludes comparisons like *wie Pilze aus dem Boden schießen* ('to spring up like mushrooms'), which are not considered in the shared task.

For all idioms that meet these conditions, we extracted all examples on the Wiktionary page that consist of at least 20 tokens. This was necessary because a lot of examples only consist of a short sentence like *Du willst mich wohl auf den Arm nehmen?* (lit. 'You want to lift me on the/your arm, right?'; fig. 'You are kidding, right?'), where it is not possible to uniquely attribute a figurative or literal use. Note that the texts in the training set, consisting of 76 tokens in average, are still much longer. The tokens in a Wiktionary example that belong to the idiom are already marked up, which facilitates converting the examples into the same format as the shared task's training data. We assigned the label FIGURATIVELY to all examples from Wiktionary, because it is untypical for dictionaries to provide examples of literal use for verbal idioms (since the literal meaning can be inferred from the meaning of the individual words).[2] In total, we collected 1,106 samples from Wiktionary.

In the next step, we collected literal counterparts for the Wiktionary samples. For this, we downloaded the 10M-sentences **News-wrt** corpus of the Leipzig Corpora Collection / Deutscher Wortschatz[3]. For collecting LITERALLY examples, we proceeded in several steps:

1. By using the German Extended Open Multilingual WordNet[4] and the WordNet interface from NLTK[5], we constructed variant forms for each idiom type in the Wiktionary collection. Here, we replaced nouns by their synonyms, hyponyms and hypernyms and kept all other words fixed. The idea is that the variant forms predominantly occur in literal meaning, while the canonical form carries the figurative meaning (Li and Sporleder, 2009).

2. We then checked for each sentence in the corpus whether it contains a variant form. Each sentence was parsed and lemmatised using spaCy[6] and a variant form was counted to appear in a sentence if all of its lemmas are directly connected to each other by dependency arcs. (This syntactic constraint drastically reduced the number of (useless) finds that we obtained with a bag-of-words approach.) Using this approach, we found 6,330 sentences.

The number of finds per type follows a Zipf distribution; Table 2 shows some examples. The most frequent type is *in Begriff sein* with 2,670 finds, which are mainly caused by the variant *in Regel sein*. *In der Regel ist etw. [...]* ('as a rule, sth. is [...]'), however, is a frequent word combination on its own and has little to do with *im Begriff sein [etw. zu tun]* (lit 'to be in the concept [to do sth.]'; fig. 'to be about [to do sth.]'). In general, types with more finds are often caused by variants that are specific constructions and thus less interesting for us.

[2] We manually checked 100 randomly selected examples. In 98 cases, we would indeed assign the label FIGURATIVELY; in 1 case, we would assign the label LITERALLY; and the remaining example did not contain a *German* idiom (but a borrowed idiom from Italian).

[3] https://wortschatz.uni-leipzig.de/en/download/German © 2021 Abteilung Automatische Sprachverarbeitung, Universität Leipzig.

[4] http://compling.hss.ntu.edu.sg/omw/summx.html

[5] https://www.nltk.org/howto/wordnet.html

[6] https://spacy.io/models/de

finds	type	example variant
2,670	*in Begriff sein*	*in Regel sein*
615	*Mann stehen*	*Person stehen*
579	*Kult sein*	*Glaube sein*
294	*in Spiel sein*	*in Finale sein*
273	*zu Bett sein*	*zu Grund sein*
209	*auf Straße gehen*	*auf Route gehen*
182	*Haus machen*	*Bau machen*
181	*in Kopf haben*	*in Politker haben*
99	*zu Sache kommen*	*zu Frage kommen*
93	*Fall setzen*	*Prozess setzen*
⋮		
10	*gegen Wand fahren*	*gegen Mauer fahren*
9	*zu Wort kommen*	*zu Begriff kommen*
9	*unter Erde bringen*	*unter Land bringen*
9	*mit Feuer spielen*	*mit Flamme spielen*
9	*in Topf werfen*	*in Kochtopf werfen*
⋮		

Table 2: Types for which variants were found in News-wrt, together with the number of finds and an example variant.

Types with less finds, on the other hand, do more frequently match variants with literal meaning. For example, the variant *mit Flamme spielen* ('to play with flame') has the same meaning as the original *mit Feuer spielen* (lit. and fig. 'to play with fire') but can only be used literally.

3. For the types with less than 100 finds, we (manually) reviewed all examples and checked whether it is possible to replace the variant with a corresponding form of the original idiom. For example *Sie spielen mit Flammen, Funken und Licht* ('They play with flames, sparks and light') is replaced by *Sie spielen mit Feuer, Funken und Licht* ('They play with fire, sparks and light'). Examples where such replacement is not possible were removed.

Unfortunately, the yield was low: In total, we could collect 36 samples from News-wrt. For these examples, we extended the context by one preceding and succeeding sentence.[7]

In a last attempt to collect some more LITERALLY examples, we used a **web** search engine to find examples for literal usages of verbal idioms in online newspaper or blog articles. Here again, we only searched idioms from the Wiktionary collection. From this web search, we could collect additional 24 examples.

Figure 1 shows the overlap (in terms of types) of the training set, our extra set and the development or test set (development and test set are interchangeable in this graphic since numbers do not change when replacing one by the other). Apparently, the extra set contributes examples for twelve types that are in the test set as well as the training set, and one type that is in the test set but not in the training set. The remaining 548 types of the extra set do neither occur in the training nor the test set.

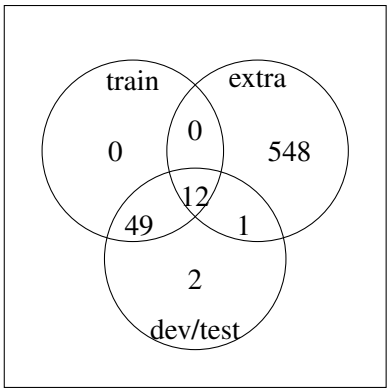

Figure 1: Overlap of training, extra and development/test set. Numbers are type counts.

3 Approach

As a baseline system, a simple linear SVM was implemented using TF-IDF features of idiom instance terms without stop words. To accommodate for the unbalanced distribution data, the model weights were adjusted. The weights are ω_l=2.92591 for the LITERALLY class, and ω_f=0.6057 for the FIGURATIVELY class, which is the balanced weights for two classes if the classes BOTH and UNDECIDABLE are disregarded (ω=0.01 each).

Furthermore, we implemented a deep learning model which employs XLM-Roberta (Conneau et al., 2020), a pre-trained neural language model that uses the transformer architecture. Specifically, we use the `xlm-roberta-base` version from the huggingface transformers library (Wolf et al., 2020). The training was performed Google Colab[8] TPUs, which allowed training runs in about 160 seconds for all epochs. The model was trained for four epochs with the same class weights as the

[7] The News-wrt corpus is scrambled but contains the URL of the website from which each sentence was crawled, so it was possible for us to receive the context from there.

[8] https://colab.research.google.com/

SVM. Furthermore, we used the following hyper parameters in Table 3:

parameter	value
sequence length	128
training batch size	64
learning rate	2e-5
L2 weight decay	0.01
LR decay	0.95
optimizer	Adam

Table 3: Hyperparameters for the XLM-RoBERTa model

For both models, we trained once on the training set as published with the shared task, once with the additional data described in Section 2, and once with the additional LITERALLY instances.

4 Results

We report the results for the baseline and XLM-RoBERTa systems trained on the training set provided by the shared task organizers and with extended data in Table 4. Models with all additional training data are reported under *SVM ext. all* and *XLM-RoBERTa ext. all*. On the other hand, *SVM ext.* and *XLM-RoBERTa ext.* are the models that were trained using only the additional LITERALLY instances.

In Table 5, we compare the results of all four models according to their performance on the LITERALLY class for seen and unseen idiom instances on the test data.

We can see that the baseline model slightly improves with the additional training data. However, the XLM-RoBERTa model trained without the additional data outperforms the model with the extended data for both classes, see Table 4 and has a higher f1-score for both seen and unseen data, see Table 5.

Out of 265 instances of the LITERALLY class in the test set, the XLM-RoBERTa model mispredicted 25 cases as FIGURATIVE, which are listed in Table 6. The Shared Task organizers (Ehren et al., 2020) present idiomaticity rates (IR) for the seen idioms in the test set, which can be found the IR column. Unfortunately, the idiom with the highest ratio of false negatives has no idiomaticity rate reported. We can observe that the examples with a lower idiomaticity rate seem to be more likely to be misclassified. However, for a thorough analysis, an higher number of total instances of idioms and the idiomaticity rate of the unseen examples would be needed.

	precision	recall	f1-score
SVM			
literally	0.6225	0.4811	0.5427
figuratively	0.8879	0.9390	0.9127
SVM ext.			
literally	0.6197	0.5000	**0.5535**
figuratively	0.8910	0.9357	0.9128
SVM ext. all			
literally	0.6070	0.4600	0.5236
figuratively	0.8870	0.9386	0.9121
XLM-RoBERTa			
literally	0.6523	0.9167	**0.7622**
figuratively	0.9776	0.8995	0.9369
XLM-RoBERTa ext.			
literally	0.6158	0.8864	0.7267
figuratively	0.9711	0.8863	0.9268
XLM-RoBERTa ext. all			
literally	0.5874	0.9167	0.7160
figuratively	0.9777	0.8666	0.9188

Table 4: Results of the baseline and XLM-RoBERTa systems with the extended training data on the development set

Model	f1-score	f1-unseen
SVM	0.5696	0.4000
SVM ext.	0.5776	0.4000
RoBERTa	0.7619	0.7381
RoBERTa ext.	0.7365	0.6728

Table 5: Results of the systems on the test data for known and unknown VIDs

Idiom	#FN	Total	IR
Frucht tragen	6	20	-
vor Tür stehen	5	27	68.17
auf Tisch liegen	3	28	71.29
Fäden ziehen	2	4	89.49
mit Feuer spielen	2	9	-
Korb bekommen	2	12	-
auf Straße stehen	1	12	87.00
Brücke bauen	1	23	92.45
in Wasser fallen	1	10	90.15
in Schatten stehen	1	1	100.00
über Bord gehen	1	9	86.30

Table 6: False negatives of the LITERALLY class by idiom (test set) and total number of LITERALLY instances in test set

Apart from this slight tendency, there is no clear pattern which idioms are more likely to be falsely predicted. Neither prepositions at the beginning of the idiom (14/25) nor total number of literal instances in the data set seem to be a direct cause for false negatives.

5 Conclusion

We presented our team's submission to the Shared Task on the Disambiguation of German Verbal Idioms at KONVENS 2021. Our best performing model, an XLM-RoBERTa model achieved 0.7622 f1-Score. While the model does not benefit from additional training data as we had hoped, it still outperforms the baseline and all submissions to the shared task by other teams. We believe that the carefully crafted, reviewed and cleaned additional data could provide valuable data in addition to the data set used in the shared task.

A conclusive study of the relationship between idiomaticity rate and performance of disambiguation systems would be an interesting objective for further study.

In conclusion, we learned that in order to predict literal use of German verbal idioms an elaborate and rather "mighty" model such as XLM-RoBERTa performs well, although one might say, we cracked a walnut with a sledgehammer, or as the Germans say: *Wir haben mit Kanonen auf Spatzen geschoßen* (lit. 'we shot sparrows with cannons').

References

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.

Rafael Ehren, Timm Lichte, Laura Kallmeyer, and Jakub Waszczuk. 2020. Supervised disambiguation of German verbal idioms with a BiLSTM architecture. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 211–220, Online. Association for Computational Linguistics.

Linlin Li and Caroline Sporleder. 2009. Classifier combination for contextual idiom detection without labelled data. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 315–323, Singapore. Association for Computational Linguistics.

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.