An Analysis of Attention in German Verbal Idiom Disambiguation

Rafael Ehren[1], **Laura Kallmeyer**[1], **Timm Lichte**[2]
[1]Heinrich Heine University, [2]University of Tübingen
{ehren,kallmeyer}@phil.hhu.de, timm.lichte@uni-tuebingen.de

Abstract

In this paper we examine a BiLSTM architecture for disambiguating verbal potentially idiomatic expressions (PIEs) as to whether they are used in a literal or an idiomatic reading with respect to explainability of its decisions. Concretely, we extend the BiLSTM with an additional attention mechanism and track the elements that get the highest attention. The goal is to better understand which parts of an input sentence are particularly discriminative for the classifier's decision, based on the assumption that these elements receive a higher attention than others. In particular, we investigate POS tags and dependency relations to PIE verbs for the tokens with the maximal attention. It turns out that the elements with maximal attention are oftentimes nouns that are the subjects of the PIE verb. For longer sentences however (i.e., sentences containing, among others, more modifiers), the highest attention word often stands in a modifying relation to the PIE components. This is particularly frequent for PIEs classified as literal. Our study shows that an attention mechanism can contribute to the explainability of classification decisions that depend on specific cues in the sentential context, as it is the case for PIE disambiguation.

Keywords: verbal idiomatic multi-word expressions, attention models, explainable AI

1. Introduction

Due to the success of the Transformer architecture (Vaswani et al., 2017), attention is one of the most popular concepts in Deep Learning right now. In NLP, BERT-based (Devlin et al., 2019) architectures are so dominant, that it seems to have given rise to the new field of 'BERTology' (Rogers et al., 2020; Søgaard, 2021), where researchers try to explore, what BERT learns about language. But it is not only the performance, which makes attention so popular, but also the fact that it gives us a certain degree of explainability, as attention weights potentially reveal what influences a model the most during a decision. However, it is currently the subject of lively debate how great this potential actually is (cf. Section 2).

In this work, we use attention in order to gain some insights into what contextualizing deep learning architectures are capable of learning when performing the task of disambiguating potentially idiomatic expressions (PIEs). PIE disambiguation is a subtask of multi word expression (MWE) identification. PIEs are potentially idiomatic, i.e., they can have a literal or an idiomatic reading like *rock the boat* ('cause trouble'):

(1) If you want that promotion, you should stop rocking the boat. IDIOMATIC

(2) They rocked the boat and fell into the freezing cold river. LITERAL

Example (1) shows a sentence containing an idiomatic usage of the PIE type, i.e. an instance of the verbal idiom (VID) type, while (2) contains an instance of its literal counterpart[1]. PIEs are challenging for NLP applications, because it is not enough to map a string to a certain VID type. To correctly disambiguate a PIE instance we have to take the context into account as well as its form, since VIDs are often subject to morhposyntactic restrictions (e.g. *kick the bucket* is not readily passivisable: **the bucket was kicked*).

In this paper, we use an established architecture for PIE disambiguation in German, based on Ehren et al. (2020), and investigate which elements of the sentential context of a PIE are crucial for deciding whether it is literal or not. More concretely, we investigate syntactic features and relations to the PIE components of those elements that are particularly indicative for literalness and idiomaticity. To this end, we propose an attention-based architecture capable of revealing which part of the context has the strongest influence on the model's classification decisions. More concretely, we stack an attention mechanism on top of the BiLSTM architecture proposed by Ehren et al. (2020) (cf. Section 4). Our architecture is applied to German verbal idioms, using the data from Ehren et al. (2020) (cf. Section 3). We opted for the former architecture instead of a BERT-based one for the sake of simplicity, comparability and greater transparency.

Our results, presented in Section 7, support the view that attention can be leveraged to make neural-network models more "explainable", as we can statistically corroborate our impression that the attention model often puts its focus on tokens that seem to be most crucial also for the human classifier. At the same time, the difficulties of the classifier with the peculiarities of the minority class becomes evident. To our knowledge, this is the first study of its kind, particularly in the area of idiom identification.

[1]The term PIE was coined by Haagsma et al. (2019) and it allows to encompass the literal and idiomatic usage at the same time.

Proceedings of the 18th Workshop on Multiword Expressions (MWE 2022) n@LREC2022, pages 16–25

2. Related Work

Attention-based models, especially BERT, have been used in the task of PIE classification (as well as many other NLP tasks) with considerable success, reaching first places in shared tasks (Taslimipoor et al., 2020; Pannach and Dönicke, 2021) or state-of-the-art results on well established data sets (Fakharian and Cook, 2021). Following the success in this and other areas of NLP, an interest in the more fine grained representational properties of these models has grown.

One way to shed more light on these models is to examine the resulting embeddings using cosine similarity. This is, for example, done in Garcia et al. (2021) for investigating the representation of compositionality in nominal compounds. Looking at pretrained embeddings from several both contextualizing and static models, they compare embeddings of compounds with the embeddings of their components, synonyms, and contexts by means of cosine similarity and find that pretrained contextualized models often do not distinguish between compositional and idiomatic compounds.

Another approach that has recently attracted a great deal of interest is to use the attention scores in attention-based models such as BERT, and to analyse the focus of attention when the model is classifying input in a certain way. An early example of such an analysis was already given by Bahdanau et al. (2016) who were the first to apply attention to a machine translation task, and who employed two-dimensional attentional heat maps to visualize the "non-monotonic" alignments between tokens of source and target language. Meanwhile, there are powerful interactive tools such as BertViz (Vig, 2019) to vizualize the attention scores of different heads and layers. At the same time, however, there is an ongoing discussion to what extent attention scores are actually useful to explain the decisions of contextualizing models (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Bastings and Filippova, 2020; Søgaard, 2021). For example, it has been claimed by Jain and Wallace (2019) that "Attention is not Explanation". In a series of experiments on binary text classification and question answering, using BiLSTMs coupled with Bahdanau Attention, they found only a weak correlation between attention weights and other, gradient-based measures of feature importance. Furthermore, they were able to find attention distributions very different from the learned ones, which nevertheless yielded nearly identical prediction scores. From this, they conclude that attention does not provide "faithful" explanations of a model's decisions. Wiegreffe and Pinter (2019) reject the assumption that an attention distribution needs to be *exclusive* to serve as explanation. In addition, they show that even when adversarial attention distributions can be found, they do not perform as well on a simple diagnostic as their learned counterparts. They conclude that explainability depends on the definition and distinguish between plausible and faithful explanations, with the former not being invalidated by the work of Jain and Wallace (2019). We agree with Wiegreffe and Pinter (2019) that exclusivity is not a prerequisite in order for an attention distribution to serve as plausible explanation. Furthermore, like the two former works we will also use a one-layered BiLSTM as an encoder, coupled with Bahdanau attention, since Wiegreffe and Pinter (2019) established that the hidden states can still act as faithful representations of the input tokens, which is very important as we want to make claims about the influence of the different inputs. It is not clear, if this also holds for a very deep encoder like a BERT-based one. In this work, we will be contributing to the question of the usefulness of attention scores by applying a statistical and introspective analysis of the main attention to the classification of PIEs.

3. Data

We perform our experiments on the COLF-VID 1.0 (COrpus of Literal and Figurative meanings of Verbal IDioms) data set (Ehren et al., 2020), which consists of 6985 sentences drawn from newspaper texts with examples of 34 German VID types. Every instance in the corpus is annotated with one of the four labels IDIOMATIC, LITERAL, UNDECIDABLE or BOTH. Only 0.59% of the instances are given one of the latter two labels, so basically we are dealing with a binary classification task. The distribution of the remaining two labels is imbalanced as 77.55% of the instances are labeled as idiomatic, while only 21.86% are judged to be literal. An example from COLF-VID 1.0 is shown in (3):

(3) Bundesbahn will die **Notbremse** **ziehen**.
Federal railway wants the **emergency brake** **pull**.
'Federal railway wants to pull the emergency brake.'

It shows a usage case for the VID *die Notbremse ziehen* ('pull the emergency brake'⇒'put an immediate hold on something') which is labeled as IDIOMATIC.

The data is split following Ehren et al. (2020): 70% of the data are used for training, while 15% are used for the dev and the test set, respectively. Since the number of instances per PIE types in COLF-VID is highly skewed, we perform a balanced split, i.e. every split contains the same ratio of instances per PIE type.

There exist a variety of similar PIE corpora that would in principle be suitable for our proposed attention architecture, for example the MAGPIE corpus (Haagsma et al., 2020). The main reason we choose COLF-VID 1.0 is its size and relatively low idiomaticity rate, and the fact that it has been used in Ehren et al. (2020), which our attention architecture builds on. We describe our architecture in the next section.[2]

[2]Another corpus of verbal PIEs, which contains COLF-

4. System

Our system is based on the BiLSTM+MLP classifier by Ehren et al. (2020) enhanced with an attention mechanism similar to the one in Bahdanau et al. (2016). Figure 1 shows the overall architecture together with an example for the input (4):

(4) Das Konzert **fiel ins Wasser**.
The conert **fell into the Water**.
'The concert was cancelled.'

In a first step shown at he bottom of Figure 1, the pretrained embeddings of the input tokens are fed into a BiLSTM. The concatenated outputs of the forward and backward LSTMs give us the contextualized version of the input embeddings, which ideally should contain information about the relevant preceding and succeeding elements in the token sequence. In Ehren et al. (2020), the contextualized embeddings are then fed into a multilayer perceptron (MLP) to conduct PIE classification. However, in our model, we add an attention mechanism between the BiLSTM and the MLP.

When talking about attention mechanisms, the terms *keys*, *values* and *query* – which all denote vectors – play an important role. We can think of the query as the vector representation of the question what the model should pay attention to, while the keys are the potential candidates for receiving this attention. Since our aim is to explore which tokens in the input sequence the model focuses on the most during classification, it makes sense to use the their contextualized embeddings. Keys and values are the same in our case. The answer what should function as the query is less obvious as there exist numerous options. Because the PIE instance is the anchor point for every classification decision, we choose the average of the pretrained embeddings of the PIE's components. Now we can compute the attention scores based on the query and the keys. Given a query $q \in \mathbb{R}^q$ and a key $k_i \in \mathbb{R}^k$ we leverage the following scoring function taken from Bahdanau et al. (2016):

$$score(q, k_i) = w_v^\top \tanh(W_q q + W_k k_i) \quad (1)$$

Here, k_i is a key, and $W_q \in \mathbb{R}^{h \times q}$ and $W_k \in \mathbb{R}^{h \times k}$ represent linear transformations mapping k and q into the same space before they are added together[3]. Then, the resulting vector is put through the $tanh$ function and is multiplied with $w_v^\top$, so we receive a single score. After we computed the attention score for every key k_i we apply *softmax* in order to obtain a probability distribution $a_{0:n}$ of attention weights over all input tokens.

With $a_{0:n}$, we compute the weighted average for the contextualized embeddings $v_{0:n}$, which gives us the context vector c that represent the context of a PIE instance:

$$c = \sum_{i=0}^{n} a_i v_i \quad (2)$$

Note that all contextualized embeddings are included, even the ones representing the PIE components, although they do not really belong to the context, but form the target expression itself. One could exclude them by setting their scores to $-\infty$, which would result in their corresponding attention weights being set to zero when fed into the softmax function (as done with the padding tokens). But as addressed earlier, it might not only be the context providing clues on the correct reading, but also the PIE constituents themselves by exhibiting morphosyntactic flexibility atypical for the respective VID.

Finally, we concatenate c with q and feed it into a MLP to compute the scores for the four classes. What we expect in this example is that the contextualized representation for the token *Konzert* receives the highest attention and thus influences the context vector the most, because it is the only token in the sentence that provides information on the correct reading of the PIE instance.

5. Disambiguation experiments

Using the same hyperparameters as Ehren et al. (2020), we train our model for 30 epochs with a batch size of 32 and employ fastText embeddings (Bojanowski et al., 2016) with 300 dimensions as input. The hidden layers of the LSTMs are of size 100 which give us contextualized vectors of size 200 after concatenation. Consequently, the context vector has the same dimensionality. For the query vector, the centroid of input embeddings is used, and its concatenation with the context vector results in an input layer of size 500 for the MLP, which has one hidden layer with 100 neurons. For optimization we use cross-entropy loss and the Adam (Kingma and Ba, 2014) variant of the SGD algorithm. The implementation can be found on GitHub[4].

Table 1 shows the results on the validation and the test. We report the weighted macro average to account for the stark imbalance in classes. Since we use the same model and data set as Ehren et al. (2020), it makes sense to compare results to those achieved by the base model[5]. To our surprise, the attention model performs slightly worse than the base model with an F1 score of 87.66 against 87.99 on the validation set and 86.89 against 87.83 on the test set.

We suspect that the reason for the decrease in performance is that, by adding the attention mechanism, we introduce an additional 60.000 parameters in the form of the two weight matrices W_q and W_k (cf. Equation 1),

VID, was used in a recent shared task at KONVENS (Ehren et al., 2021).

[3]Note that k and q might already be in the same space if the contextualizations and embeddings have the same dimensionality.

[4]https://github.com/rafehr/PIE-attention

[5]More precisely, to the results with the model using fastText embeddings. Ehren et al. (2020) also employ word2vec and ELMo.

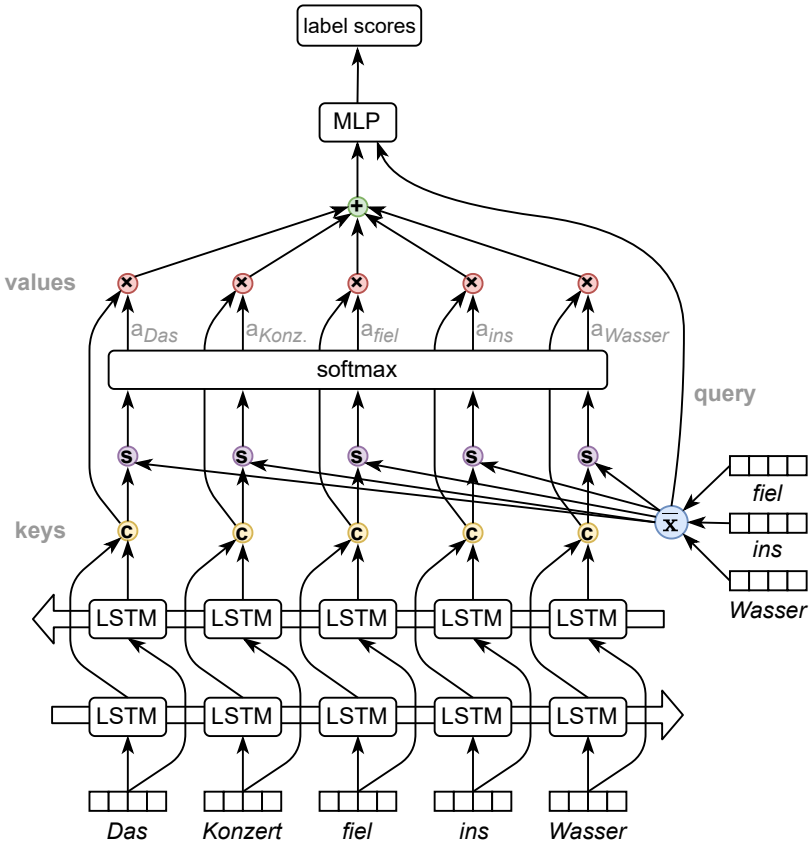

Figure 1: Architecture of the attention model.

		Weighted macro average		
Model	**Split**	**Pre**	**Rec**	**F1**
Majority baseline	Val	56.78	75.32	64.75
	Test	59.22	76.95	66.93
Ehren et al. +fastText	Val	87.86	88.14	87.99
	Test	87.45	88.29	87.83
This work	Val	87.44	87.88	87.66
	Test	86.83	86.89	86.85

Table 1: Evaluation results of the attention model on the COLF-VID 1.0 data set and comparison to baseline models

which make up the attention scoring function and were both of size 100×300. For training a model with that many parameters, our data set might be too small. This is supported by the fact that other parameter increasing measures during hyperparameter tuning like an enlargement of hidden layer size or hidden layer number all result in (far) worse performance. We refrain from more extensive hyper parameter tuning, since our focus is not on performance but on using the attention mechanism for purposes of explainability.

6. Extracting properties of tokens that receive a high attention

Our main goal is to uncover which parts of the input the model pays most attention to and what this might tell us about what it is learning in this kind of task. Therefore our architecture is designed in a way that attention scores are expected to have considerable influence on the classifier's decision: Everything the MLP sees at the end is a context vector which is composed of contextualized fastText embeddings weighted by their respective attention score.

We are particularly interested in the *maximum attention token* (MAT) of PIE contexts, i.e., the token that receives the highest attention, and we inspect the following properties of the MAT: (i) its attention weight, (ii) its POS tag, and (iii) the label of the first arc on the dependency path between the verb component (respectively the noun component) of the PIE and the MAT.

In order to gather this information, we parse the sentences using the NLP library spaCy[6], which gives a labeled dependency tree for every sentence. The POS tagging is conducted with the TreeTagger (Schmid, 1999), which uses the STTS tag set. We group the STTS POS tags into four general categories: noun (NN, NE), verb (VV*, VA*, VM*), adjective (ADJD, ADJA), and other. Note that we use the dependency

[6] https://spacy.io/

parses and POS tags only for the attention statistics; the PIE disambiguation classifier does not use syntactic information but acts solely on surface tokens.

Concerning the dependency labels, there are obviously cases where we do not have a direct arc between the respective PIE component and the MAT, but we always have a dependency path, provided parsing was successful. We assume that the label of the first arc on this path, starting from the PIE component, is a good choice for characterizing the relevant aspect of the dependency relationship between the two words, since it indicates the relation between the PIE component and the MAT including its dependency context. For illustration, consider Figure 2, which shows an idiomatic usage of *in der Luft hängen* ('hang in the air'⇒'be present').[7] Components of the PIE are bold, the MAT

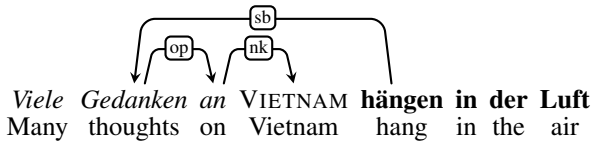

Figure 2: Subject (SB) relation between the verb and the noun phrase containing the MAT *Vietnam*

in capital letters and the rest of the sentence is in italic. There is no direct arc from the PIE verb to the MAT, but there is a path from the subject of the PIE verb to the MAT (VIETNAM), since the latter is part of a PP that modifies the subject. Thus, since the MAT is part of the subject NP, the system pays attention to some property of the subject. Such examples motivate our choice to register the first label (here SB) on the path from PIE component to MAT.

There is one more peculiarity with regard to how we register dependency relations. Very often – in 20.38% of the cases to be exact – the first arc in the (undirected) path from the PIE verb to the MAT is labeled OC for *object clause*, see for example Figure 3. Here the head

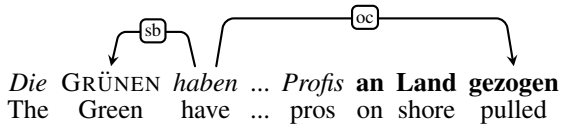

Figure 3: OC (object clause) relation between the PIE verb and the finite auxiliary verb

of the PIE verb is the auxiliary *haben* which in turn governs the subject. In such cases, we disregard the OC relations and register the label first label on the path from PIE component to MAT that is not OC (SB in this case).

[7] Another meaning of *in der Luft hängen* is 'to be uncertain'.

	FIG	LIT	overall
average MaxAttn	0.52	0.46	0.51
STD	0.2	0.18	0.2
MaxAttn on PIE verb (%)	1.23	2.92	1.6
MaxAttn on PIE noun (%)	6.51	13.75	8.11
MaxAttn on noun (%)	82.06	71.25	79.53
MaxAttn on adjective (%)	9.21	15.00	10.66
MaxAttn on verb (%)	3.56	7.5	4.43
MaxAttn on other (%)	5.16	6.25	5.38
MaxAttn on sb (%)	39.8	17.08	34.62
MaxAttn on mo (%)	25.8	41.67	29.43

Table 2: Selection of global attention statistics

7. Attention statistics

We collect the attention scores on the test set and compute statistics individually for instances where the system predicts the label FIGURATIVE (FIG) and for instances where the label LITERAL (LIT) is predicted.[8] Finally, we also perform an ablation experiment by replacing noun MATs with pronouns, in order to assess whether the system pays attention rather to grammatical functions or to semantic properties of lexical items.

7.1. Global attention statistics

Table 2 shows a selection of the global attention statistics. The first column contains the numbers for FIG, the second for LIT, and the last for FIG and LIT combined.

First, not surprisingly, for both classes, LIT and FIG, the model focuses more on content words than on function words, since the vast majority of MATs have POS tags of nouns and adjectives. However, there is a considerable difference between the two classes: LIT has a much larger preference for (adverbial/predicative) adjectives than FIG (15 % vs. 9.21 %) and a lower preference for nouns (71.25 % vs. 92.06 %).

Concerning dependency relations, in FIG sentences, subjects are more likely to contain a MAT compared to LIT. The reason might be that for the verb (without the PIE context), the literal reading is much more frequent, and in idiomatic readings, we might have subjects whose semantic properties are in contradiction to the semantic features that subjects of the literal reading usually have. Put differently, the choice of the subject filler is more marked in figurative readings than in literal ones.

This is in line with our experience when annotating PIEs, where selectional preference violation was identified as one of the key factors to inform the decision whether a PIE instance was idiomatic. The following example shows such a violation:

[8] The other two labels are barely predicted at all, so we do not include those in the statistics.

(5) But the **White House** is **playing with fire** by not complying here [...].[9]

Here the subject is an institution instead of the animate agent we would expect with the verb *play*, thus revealing the idiomatic reading.

Another salient observation is the magnitude of the attention given to the MAT by the system: the mean attention is 0.51 with a standard deviation of 0.2. This indicates that the attention is rather not distributed between multiple tokens. On the contrary, the model seems to pick one target that clearly stands out in terms of attention score, since, on average, MaxAttention differs considerably from the second highest attention score. The minority class LIT has a smaller MaxAttention than the majority class FIG, which seems to reflect the uncertainty of the classifier and the difficulties to identify clear indicators of LIT instances.

A further noticeable difference can be observed in the ratio of cases in which the MaxAttention is on PIE elements: again this could be taken to speak for the uncertainty of the classifier regarding LIT instances; or it might be the case that morphology contributes crucial indicators by deviating from the canonical form we expect in FIG instances. Note that fastText embeddings also take morphological features into account by virtue of the subword method. However, a manual inspection of the nominal PIE elements with MaxAttention failed to confirm that they are consistently morphologically non-canonical with respect to FIG usage. A more detailed investigation of why the model chooses a PIE element in some cases is left for future work.

7.2. Attention scores and sentence length

Since the features we investigated above can vary considerably depending on the size of the sentence, we also plotted them against sentence length, distinguishing again between FIG and LIT.

Figure 4 and Figure 5 show how the maximal, second highest and average attention (RestAttention, not counting maximal attention) scores develop with increasing sentence length. The solid line is the mean, while the area surrounding it represents the 95 % confidence interval. In both LIT and FIG, MaxAttention decreases with increasing sentence length, albeit Pearson's correlation coefficient is only weakly negative (overall -0.267 for sentences up to 30 tokens). Second highest attention and RestAttention remain rather stable, and in both LIT and FIG, the difference between MaxAttention and second highest attention seems pronounced, while in LIT the confidence interval almost overlaps in some areas, which is clearly not the case for FIG. Generally, second highest attention and RestAttention are relatively close. Again, the larger confidence area and the slightly (but not significantly) lower

Figure 4: Attention and sentence length for FIG

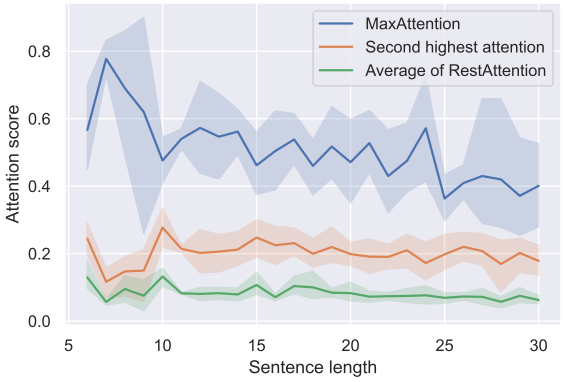

Figure 5: Attention and sentence length for LIT

MaxAttention mean for LIT seems to suggest that the classifier is struggling more to find good indicators for LIT than for FIG, regardless of the sentence size.

7.3. Syntactic features of MATs and sentence length

The development of syntactic properties of the MAT (POS tag and dependency label) is plotted against sentence length in Figure 7 for LIT and in Figure 6 for FIG. Again, we observe very different patterns in the two cases.

First, as already mentioned above in connection with Table 2, we see that MATs are more often contained in subjects (relation SB) of figurative PIEs, compared to literal PIEs; for longer sentences the difference is even more striking than the overall values from Table 2.

A second observation is that, for LIT, modifiers (relation MO) quickly become more important than subjects. Thus, for longer sentences in LIT, modifiers seem to be rather indicative for the label. And although adjectives play a larger role in LIT, especially for shorter sentences, the most frequent general POS tag for MATs is noun as can be seen from Figure 7. A manual inspection of the data suggests that nominal MATs with a modifying relation to the PIE verb are often the heads

[9] https://www.politico.com/newsletters/playbook/2019/10/08/trump-changes-the-subject-486633, accessed 04/11/2022

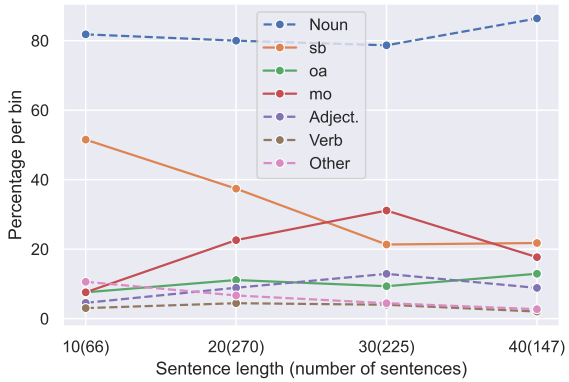

Figure 6: POS/dep. labels and sentence length for FIG

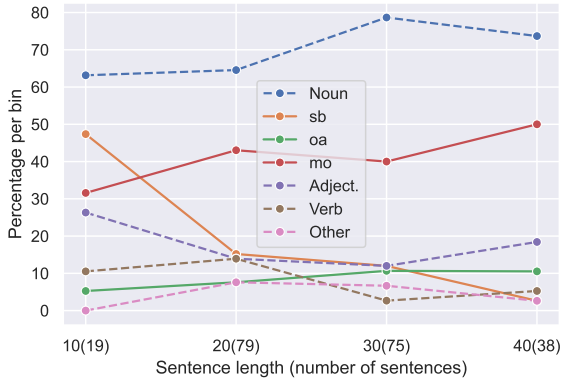

Figure 7: POS/dep. labels and sentence length for LIT

of locative PPs.

7.4. Ablation test using pronouns

The goal of replacing MAT nouns with pronouns – while taking care that the remaining sentence is still grammatical – is to test whether it is the grammatical function which the model likes to pay attention to, or rather some token in the context of the PIE by virtue of being a content word. For this, we manipulate a subset of 474 PIE instances and compute the attention statistics as done above. Because of the increasing data sparseness, we concentrate on FIG with 339 instances and compare them with the attention scores of the unmanipulated source.

The overall attention scores for the original and manipulated FIG instances are shown in Figure 8 and Figure 9) respectively. We can observe that the MaxAttention decreases, compared to the original data, but the pattern basically remains intact.

Figure 10 and Figure 11 plot the MAT's syntactic features against sentence length for the original and pronominalized FIG instances respectively. A general observation in both cases is that, after pronominalization, nominal POS tags and SB dependencies receive less attention than before; i.e., the MaxAttention does

Figure 8: Attention and sentence length for FIG before pronominalization

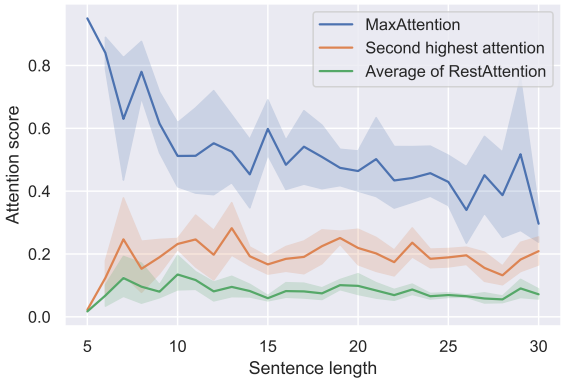

Figure 9: Attention and sentence length for FIG after pronominalization

not tend to remain on the role filled with the new pronoun (the POS tags for pronouns account for only 2.5% in total). Modifiers (MO), on the other hand, receive more frequently the highest attention, in particular for short sentences. This seems to indicate that the model pays attention to combinations of subject dependency label and content word and, in the absence of this, tends to turn to modifiers.

8. Qualitative analysis

To gain a better intuition for the attention preferences of the model, we now turn to a qualitative analysis of some of the data. We will look into examples from the perspective of an annotator in order to explore whether the systems attention falls on tokens a human would also consider important for their decision to annotate a PIE instance in a certain way. The example sentences below are equipped with a heatmap indicating the weight distribution - the higher the attention, the more intense the color.

Example (6) shows an instance of the PIE *auf dem Tisch liegen* ('lay on the table'⇒'be available/be known') with *Zahlen* ('numbers') as subject:

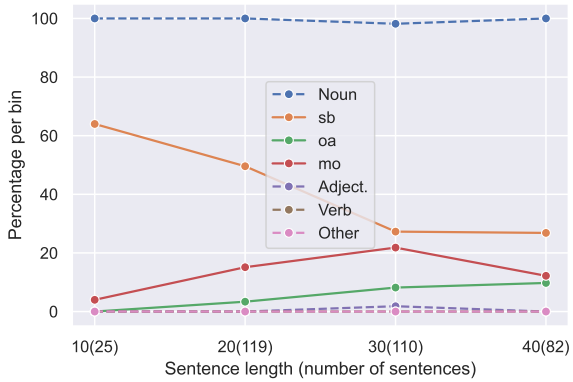

Figure 10: POS/dep. relation vs. sentence length for FIG before pronominalization

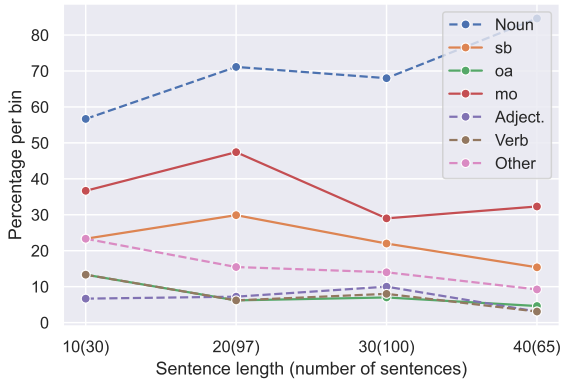

Figure 11: POS/dep. relation vs. sentence length for FIG after pronominalization

(6) *Diese Zahlen* **lagen** *am Morgen danach*
These numbers **lay** on the morning after
bereits **auf** *Erich Honeckers* **Tisch.**
already **on** Erich Honecker's **table**.
'These numbers were already reported to Erich Honecker the following morning.'

We can interpret the abstractness of the subject as an indicator for the idiomatic reading, since numbers (usually)[10] cannot be placed on a table. The model set the same focus and in four of four cases, in which *Zahlen* was the subject of *auf dem Tisch liegen*, it received the highest weight and the label *FIG* was predicted.

[10]We could of course construct a context with physical representations of numbers, but this is obviously not the case here. A bigger problem is that we can interpret it metonymically with *numbers* standing for a phyisical report lying on someone's table. But the annotators of COLF-VID did not follow this route and usually judged these type of instances to be figurative.

In (7) we have one of eight instances of the PIE *eine Brücke bauen* ('build a bridge'), where *Brücke* ('bridge') was modified with the adjective *goldene* ('golden') which gives rise to the idiomatic meaning 'give someone an easy way to retreat'.

(7) *So werden dem künftigen*
This way will be the future
Bankkunden goldene **Brücken** *bis zu*
bank customer golden **bridges** including
Zinssparen und Dispokredit **gebaut.**
interest saving and overdraft credit **built.**
'This way, golden bridges will be built for the future bank customer as far as interest savings and overdraft facilities.'

Since bridges are seldomly built from gold, the presence of the adjective is very informative to establish the correct reading. The model did pick up on that fact as *goldene* is in the top 3 of tokens with the highest attention in seven of eight cases, predicting FIG six times.

Another adjective attracting a lot of attention is *tief* ('deep'), when used adverbially with Luft holen (''take a breath'⇒'to take a break') as shown in (8).

(8) *Wer dort tief* **Luft holt,** *kann den Duft*
Who there deeply **air takes**, can the smell
des Newlands Stadium in Kapstadt
of the Newlands Stadium in Cape Town
einatmen [...] .
breathe in [...].
'If one takes a deep breath, one can breathe in the smell of the Newlands Stadium in Cape Town.'

In 9 of 12 of those cases the system gave the highest attention to *tief*, predicting the class LIT eight times. But in contrast to the examples above, it actually is not a sure sign for a literal reading, because it can just as well modify the idiomatic reading (*take a deep breath* ⇒ *take a long break*), as is represented in the test set, since 6 of the 12 instances were actually labeled as idiomatic. But since roughly 70% of instances in the training set occurring with *tief* were labeled as literal, the model reasonably predicted the label LIT.

More examples in which the model paid attention to tokens that a human annotator would also consider highly relevant for the disambiguation task can be found when examining the four literal instances of *im Blut haben* ('have in one's blood'⇒'have a predisposition for sth.') in the test set. In each of these cases, the object of the PIE, that represented a substance a person can actually have in their blood, was given the second or third highest attention (*Schadstoffe* ('pollutants'), *Cholesterinkonzentrationen* ('cholesterol concentration'), *Kokain* ('cocaine'), *Alkohol* ('alcohol')), while always predicting the correct reading.

(9) gives an example that was misclassified by the model since LIT was predicted although FIG would

have been correct.

(9) *Wer hat die größte, die schönste*
Who has the biggest, the most beautiful
Brücke gebaut?
bridge built?
'Who has established the best connection?'

However, the error is understandable; without context, a human annotator would also classify (9) as LIT, because of the attributes *größte* ('biggest') and *schönste* ('most beautiful') which modify *Brücke* ('bridge') (and which the attention model also focuses on).
Even though we could present many more of these types of examples, we of course do not claim, that our model's decisions correspond always to the way humans would decide between LIT and FIG concerning the role that the different input tokens play for this decision. There are a lot of instances to be found where the highest weights are associated with input tokens, that – from a human perspective – do not seem to be informative for the disambiguation. This is partly due to biases from training data, which distinguish of course our system from a human native speaker. But with our experiments, we were able to show two things: (1) The attention distribution is not arbitrary. This is not only supported by the statistics presented above, but also by a qualitative analysis of the data. (2) The relationship between the input and the output tends to be tangible and straightforward, i.e. a human can comprehend why the model focused on certain tokens. This is not self-evident, since with contextualizing models like a BiLSTM we cannot automatically assume that the hidden states are still faithful representations of the input tokens. It would be interesting to see whether a BERT-based encoder with its many layers would still allow for such a straightforward interpretation.

9. Conclusions

In the context of PIE disambiguation, we have provided strong evidence in support of the view that, for certain deep learning architectures, attention can be leveraged to uncover the influence of input tokens on the classifier's decision. Strikingly, regardless of classes and ablation measures, the attention model seems to pick exactly one pivotal target that clearly stands out compared to other tokens in the sentence in terms of attention scores. It would be interesting to explore, whether adversarial attention distributions in the same vein as for Jain and Wallace (2019) (cf. Section 2) can be found and, if so, which properties they would reveal compared to the one presented in this paper. Regardless of the outcome of such experiments, we would maintain that the results presented here are a valid, because plausible, explanation for the model's behaviour, since we do not agree that an attention distribution needs to be *exclusive* to serve as explanation.
Furthermore, the statistical behaviour of the studied attention model can be motivated with specific properties of the classes LIT and FIG, which differ considerably with respect to the syntactic categories that the model assigns MaxAttention to. This is even more apparent when taking sentence length into account, and also supported by an ablation test using pronominalization that we conducted. This work leaves many interesting options for future work, for example, the consideration of further linguistic features and ablation tests, crosslingual comparisons, and last but not least the comparison to other attention models such as BERT's self attention.

Acknowledgements

We thank the three anonymous reviewers for valuable comments.

10. Bibliographical References

Bahdanau, D., Cho, K., and Bengio, Y. (2016). Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473v7*.

Bastings, J. and Filippova, K. (2020). The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online. Association for Computational Linguistics.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv:1607.04606*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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.

Ehren, R., Lichte, T., Kallmeyer, L., and Waszczuk, J. (2020). Supervised Disambiguation of German Verbal Idioms with a BiLSTM Architecture. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 211–220.

Ehren, R., Lichte, T., Waszczuk, J., and Kallmeyer, L. (2021). Shared task on the disambiguation of German verbal idioms at KONVENS 2021. In *Proceedings of the Shared Task on the Disambiguation of German Verbal Idioms at KONVENS 2021*.

Fakharian, S. and Cook, P. (2021). Contextualized embeddings encode monolingual and cross-lingual knowledge of idiomaticity. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 23–32, Online. Association for Computational Linguistics.

Garcia, M., Kramer Vieira, T., Scarton, C., Idiart, M., and Villavicencio, A. (2021). Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the*

Association for Computational Linguistics: Main Volume, pages 3551–3564, Online. Association for Computational Linguistics.

Haagsma, H., Nissim, M., and Bos, J. (2019). Casting a wide net: Robust extraction of potentially idiomatic expressions. *arXiv:1911.08829v1*.

Haagsma, H., Bos, J., and Nissim, M. (2020). MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Jain, S. and Wallace, B. C. (2019). Attention is not Explanation. 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 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv:1412.6980*.

Pannach, F. and Dönicke, T. (2021). Cracking a walnut with a sledgehammer: XLM-RoBERTa for German verbal idiom disambiguation tasks. In *Proceedings of the Shared Task on the Disambiguation of German Verbal Idioms at KONVENS 2021*.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Schmid, H. (1999). Improvements in part-of-speech tagging with an application to German. In Susan Armstrong, et al., editors, *Natural Language Processing Using Very Large Corpora*, pages 13–25. Springer, Dordrecht.

Søgaard, A. (2021). *Explainable Natural Language Processing*. Number 51 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, San Rafael, CA.

Taslimipoor, S., Bahaadini, S., and Kochmar, E. (2020). MTLB-STRUCT @Parseme 2020: Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 142–148, online. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Vig, J. (2019). A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Wiegreffe, S. and Pinter, Y. (2019). Attention is not not explanation. 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 11–20, Hong Kong, China. Association for Computational Linguistics.