An Analysis of Attention in German Verbal Idiom Disambiguation

MWE Workshop 2022

Rafael Ehren[1], Laura Kallmeyer[1], Timm Lichte[2]

[1]University of Düsseldorf, [2]University of Tübingen

June 25, 2022

Introduction

- Task: Disambiguation of potentially idiomatic expressions (PIEs).
- Literal vs. idiomatic

  (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

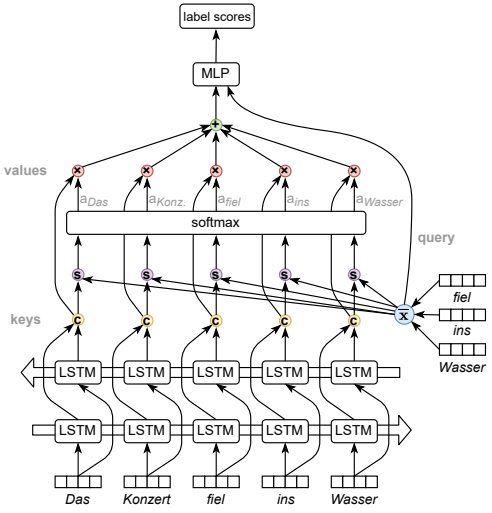
- Goal: Gain insight into which parts of the input a contextualizing neural model focuses on during classification.
- Means: Equip an existing BiLSTM model (Ehren et al. 2020) with an attention mechanism.

Related Work

- Jain & Wallace (2019): “Attention is not Explanation”
  - Experiments on binary text classification and question answering
  - Weak correlation between attention and other, gradient-based measures of feature importance
  - Found alternative (adversarial) attention distributions that resulted in the same scores
- Wiegreffe & Pinter (2019): “Attention is not not Explanation”
  - Reject the claim that an attention distribution needs to be *exclusive* to serve as explanation (*plausible* vs. faithful)
  - Show that adversarial distributions do not perform as well on a simple diagnostic as their learned counterparts

Model Architecture

- Input: FastText (Bojanowski et al. 2016) embeddings to capture morphosyntactic variation
- One-layered BiLSTM to contextualize the input
- Bahdanau attention (Bahdanau et al. 2016) on top: $score(q, k_i) = w_v^\top \tanh(W_q q + W_k k_i)$
- Keys: Contextualized embeddings; Query: Avereaged embeddings of PIE components
- Why not BERT? Embeddings contextualized by shallow BiLSTMs can still serve as "faithful" representations of the input (Wiegreffe & Pinter 2019) – unclear whether this holds for BERT with its many layers



Data

- Experiments performed on COLF-VID 1.0 (COrpus of Literal and Figurative Verbal Idioms) (Ehren et al. 2020)
- Data set consists of 6985 sentences drawn from newspaper texts with examples of 34 German PIE types
- Labels: IDIOMATIC, LITERAL, UNDECIDABLE or BOTH → Only 0.59% were labeled as one of the latter two, thus basically binary classification
- Data skewed with an idiomaticity rate of 77.55% (vs. 21.86% literal instances)

Disambiguation Results

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

$\Longrightarrow$ The performance becomes slightly worse across the board when adding the Bahdanau attention. This can be explained by the fact that more parameters have to be learned without more data being available.

Attention Statistics

- Focus on the token assigned the highest attention weight; we termed it the MAT (Maximum Attention Token)
- For the MAT, the following information was collected:
  1. its attention weight
  2. its POS tag
  3. the label of the first arc on the dependency path between the verb component (respectively the noun component) of the PIE and the MAT
- To this end, all sentences were parsed with spaCy; the POS tagging was conducted with the TreeTagger (Schmid 1999)
- Statistics were computed individually for instances where the system predicted FIG and for instances where it predicted LIT

Attention Statistics – Dependency Relations

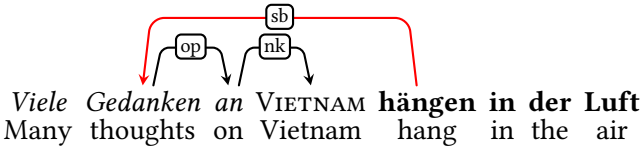


Figure: The subject relation is chosen, because it is the first arc between the PIE verb and the NP containing the MAT is chosen.

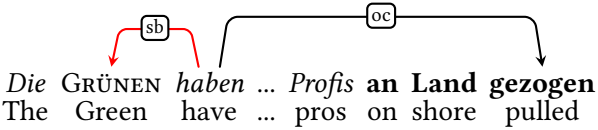


Figure: The subject relation is chosen, whereas the first arc is ignored because it is a object clause relation.

Global Attention Statistics

	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: Selection of global attention statistics

$\Longrightarrow$ The vast majority of MATs have POS tags of nouns and adjectives, while not being a component of the PIE. The difference between FIG and LIT is striking particularly with regard to dependency labels.

Attention Scores and Sentence Length

Figure: Attention and sentence length for FIG

Figure: Attention and sentence length for LIT

$\Longrightarrow$ Generally, MaxAttention decreases with increasing sentence length, while the difference between MaxAttention and second highest attention remains pronounced. As for LIT, the classifier seems to struggle more to identify a MAT.

Synt. Features of MATs and Sentence Length

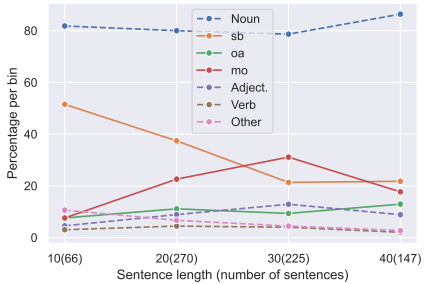


Figure: POS/dep. labels and sentence length for FIG

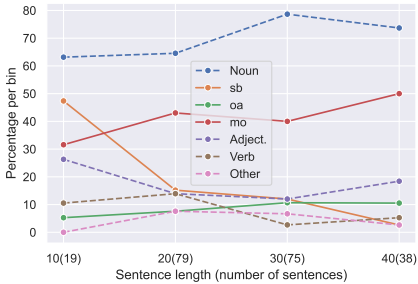


Figure: POS/dep. labels and sentence length for LIT

$\Longrightarrow$ In general, the proportion of subjects containing the MAT decreases with sentence size, while the proportion of modifiers increases. In LIT, the proportion of modifiers is much larger compared to FIG.

Ablation Test: Replace 339 MATs with Pronouns

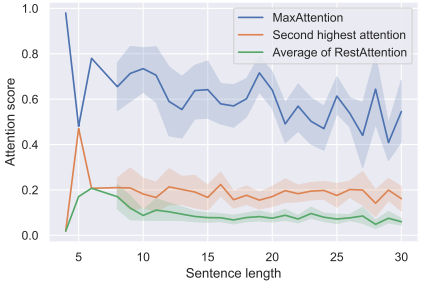


Figure: Attention and sentence length for FIG before pronominalization

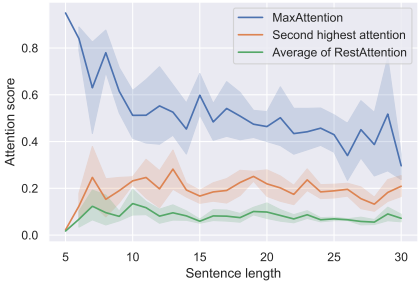


Figure: Attention and sentence length for FIG after pronominalization

$\Longrightarrow$ The MaxAttention decreases, compared to the original data, but the pattern basically remains intact.

Ablation Test: Replace 339 MATs with Pronouns

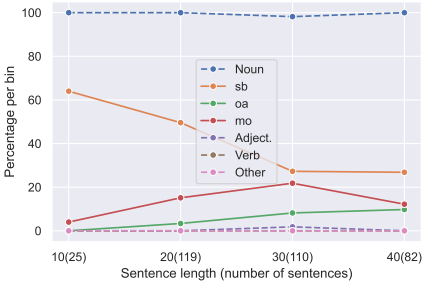


Figure: POS/dep. relation vs. sentence length for FIG before pronominalization

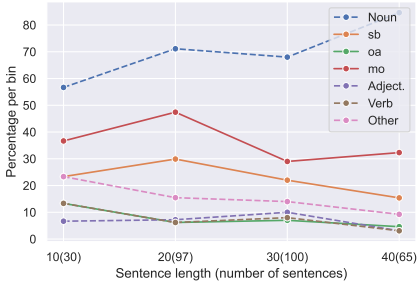


Figure: POS/dep. relation vs. sentence length for FIG after pronominalization

$\Longrightarrow$ MaxAttention does not tend to remain on the role filled with the new pronoun. Rather, MaxAttention seems to shift to other content words in its context, in particular modifiers.

Qualitative Analysis

- Does the classifier pay attention to the same tokens a human annotator would? There are quite a few examples, where it seems that it does:

(3) *So* *werden* *dem* *künftigen* *Bankkunden* *goldene* **Brücken** *bis* *zu*
This way will be the future 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.'

→ In 7 of 8 cases, the adjective *golden* was in the top three of the tokens with the highest attention, when combined with *Brücke bauen* ('build bridge')

Conclusion/Future Work

- Conclusion:
  - Strong evidence for the view that “Attention is not not explanation” (at least for PIE disambiguation)
  - Strikingly, the attention model tends to pick one pivotal item it focuses on, instead of distributing the attention over many tokens.
  - Considerable differences with regard to the two classes FIG and LIT
- Future work:
  - Explore whether adversarial attention distributions can be found and which properties they have compared to the one presented above.
  - Contextualize with BERT instead of a BiLSTM
  - Examine other languages

References

Bahdanau, Dzmitry, Kyunghyun Cho & Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473v7* https://arxiv.org/abs/1409.0473v7.

Bojanowski, Piotr, Edouard Grave, Armand Joulin & Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv:1607.04606* .

Ehren, Rafael, Timm Lichte, Laura Kallmeyer & Jakub Waszczuk. 2020. Supervised Disambiguation of German Verbal Idioms with a BiLSTM Architecture. In *Proceedings of the Second Workshop on Figurative Language Processing*, 211–220.

Jain, Sarthak & Byron C. Wallace. 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)*, 3543–3556. Minneapolis, Minnesota: Association for Computational Linguistics. doi:10.18653/v1/N19-1357. https://aclanthology.org/N19-1357.

Schmid, Helmut. 1999. Improvements in part-of-speech tagging with an application to German. In Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann & David Yarowsky (eds.), *Natural language processing using very large corpora*, 13–25. Dordrecht: Springer. doi:10.1007/978-94-017-2390-9_2.

Wiegreffe, Sarah & Yuval Pinter. 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)*, 11–20. Hong Kong, China: Association for Computational Linguistics. doi:10.18653/v1/D19-1002. https://aclanthology.org/D19-1002.