Can Yes–No Question-Answering Models be Useful for Few-Shot Metaphor Detection?

Lena Dankin
School of Computer Science
Tel Aviv University
lenadank@tau.ac.il

Kfir Bar
School of Computer Science
Reichman University
kfir.bar@post.runi.ac.il

Nachum Dershowitz
School of Computer Science
Tel Aviv University
nachum@tau.ac.il

Abstract

Metaphor detection has been a challenging task in the NLP domain both before and after the emergence of transformer-based language models. The difficulty lies in subtle semantic nuances that are required to be able to detect metaphor and in the scarcity of labeled data. We explore few-shot setups for metaphor detection, and also introduce new question-answering data that can enhance classifiers that are trained on a small amount of data. We formulate the classification task as a question-answering one, and train a question-answering model. We perform extensive experiments for few shot on several architectures and report the results of several strong baselines. Thus, the answer to the question posed in the title is a definite "Yes!"

1 Introduction

In the past year, pretrained language models established themselves as the foundation for state-of-the-art solutions for most of the common NLP tasks. Usually, one should fine tune a model on a dataset specific to her task and domain so as to achieve high performance, and this requires labeled data, which is not always available in the necessary quantity. In the past few years, a large body of work has been dedicated to transfer learning between domains and models (Alyafeai et al., 2020), and application of models trained on one task to another task by prompting (Brown et al., 2020; Schick and Schütze, 2021). These techniques reduce the amount of training data needed for a specific task, and enable the sharing of semantic knowledge between models.

Metaphor detection is a highly challenging task in the NLP domain. It relies on word level, delicate semantics that are not trivial even for humans, and, thus, even though pretrained language models do encode some metaphoric information (Aghazadeh et al., 2022), the task is not considered solved. As for languages other than English – high quality language models are already often available (Seker et al., 2021; Antoun et al., 2020), but metaphor detection without appropriate labeled data is very difficult (Schneider et al., 2022), and this is why few-shot is a relevant scenario to study.

As Su et al. (2020) suggest, metaphor detection can be viewed as a reading-comprehension task where one needs to answer a question whether a specific word is metaphoric or literal in the context of a given sentence. They formatted metaphor detection as a classification task of the full sentence (global context), the word in question and a short sentence fragment that contains this query word (local context). The texts, along with POS tags of each word, are fed into the classifier to obtain a prediction. Similar to the vast majority of classification tasks, this classifier is expected to learn how to identify metaphors based on the labels it is provided during training, but the input itself does not suggest that the task is regarding metaphor.

We take the reading-comprehension approach further in two respects: First, we experiment with several phrasings of explicit natural-language questions about whether the query token is metaphoric within the context of the sentence. Thus we employ the capability of large language models to understand delicate semantics (at least up to some point) by querying the models directly. Second, we design our classifier with a backbone of a yes–no question-answering model. Given the context sentence, we ask explicitly, "Is the word in question metaphoric in this context?" We evaluate our

Proceedings of the 3rd Workshop on Figurative Language Processing (FLP), pages 125 - 130
December 8, 2022 ©2022 Association for Computational Linguistics

model in a few-shot scenario and compare it to several baselines.

2 Related Work

Over the past few years, as in other fields in NLP, transformer-based architectures have dominated the models for metaphor detection. Leong et al. (2020) report the results of the 2020 shared task, and can be referred to for prior models that are not transformer based.

DeepMet (Su et al., 2020), the highest-scoring system in that shared task, transforms metaphor detection into a reading comprehension task, querying for the label of each token given its context in the sentence. The classification model is a siamese network that encodes two contexts for the token – the entire sentence and the sentence fragment wherein the token is located. The model is also fed with the POS tag of the token in question.

MelBERT (Choi et al., 2021) is a transformer-based model that applies two theoretical concepts of metaphor identification: (1) A metaphor's literal meaning is different from its metaphorical meaning in the sentence. (2) The metaphor is unusual in the context of the sentence. MrBERT (Song et al., 2021) employs a similar architecture to MelBERT, adding the encoded grammatical local context of the query token.

Few-shot learning refers to learning from a small number of training examples. One few-shot technique for NLP is pattern exploiting training (PET) (Schick and Schütze, 2021) over the RoBERTa architecture. PET, requiring task-specific unlabeled data, uses natural language patterns to represent the inputs as cloze style questions. Answers are then filled in by the predictions of the language model. ADAPET (Tam et al., 2021) extends PET by providing denser supervision during fine-tuning, outperforming PET without the need of unlabeled data.

GPT-3 (Brown et al., 2020) takes few-shot abilities forward and demonstrates strong performance without directly fine-tuning on task-specific data. Instead, in the few-shot scenario, at inference time it is presented with a few labeled instances as a part of the query.

3 Metaphor Detection Model

Metaphor detection can be regarded as a token-classification task within a sentence. The word in question in a given sentence can be classified either as metaphoric or literal.

In this work, we experiment with the formulation of metaphor detection as a yes–no question answering (QA) task with two concatenated inputs: a *question* and a *passage*, that is, a text segment to which the question refers. For each word in question, we suggest several different constructions of *questions* and *passages*. These formulations are shown in Table 1 and are referred to as f1–f3. We add f4 to assess the contribution of a question-like phrasing.

Our suggested architecture for metaphor detection is presented in Figure. 1. We begin by fine-tuning a RoBERTa base model (Liu et al., 2019) on QA data (see Section 4.1). Next, this model is fine-tuned on different sizes of metaphor data, phrased as questions.

The results are compared to the RoBERTa base model and to DeepMet. Since we are aiming to analyse the advantages of the QA model in a few-shot scenario, rather than to outperform the state of the art, our baseline models are ones that are similar in terms of architecture and additional resources. Training on the entire VUA dataset we are experimenting with, the RoBERTa baseline achieves the F1 score of 71.4, while MelBERT, the current state of the art, attains an F1 of 72.3.

4 Data

4.1 Yes–No Question-Answering Datasets

BoolQ. BoolQ (Clark et al., 2019) is a reading comprehension dataset comprised of 13K yes–no questions on various topics, each question relates to a different passage. The train split consist of 9.4K instances, with a ratio of 0.62 positive:negative labels.

WordNet. We utilize WordNet (Fellbaum, 1998) to extract yes–no questions to train a question answering model. WordNet curates a large collection of English lexemes, along with their different senses and different usage examples for each. When the different meanings are completely unrelated (like the word *bank* used for a financial institution or for sloping land), we rely on the context to determine the right meaning. This is somewhat related to the task of metaphor detection due to the fact that the model needs to address the alternative meanings a word may have.

For each *word* and *sentence example*, we con-

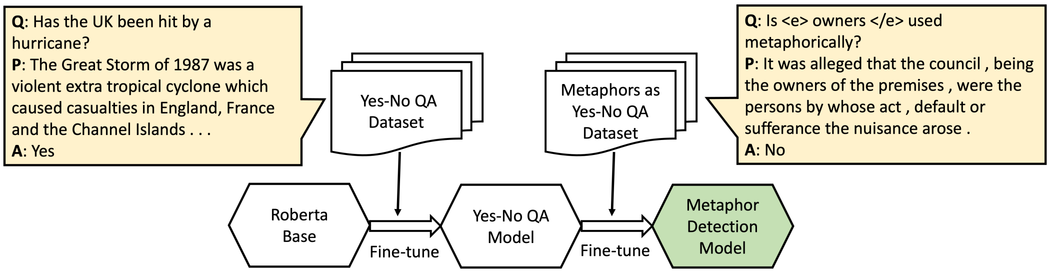

Q: Has the UK been hit by a hurricane?
P: The Great Storm of 1987 was a violent extra tropical cyclone which caused casualties in England, France and the Channel Islands . . .
A: Yes
Yes-No QA Dataset
Metaphors as Yes-No QA Dataset
Q: Is <e> owners </e> used metaphorically?
P: It was alleged that the council , being the owners of the premises , were the persons by whose act , default or sufferance the nuisance arose .
A: No
Roberta Base
Fine-tune
Yes-No QA Model
Fine-tune
Metaphor Detection Model

Figure 1: Our suggested metaphor detection model is fine-tuned on top of a yes–no question answering model.

	Question	Passage
f1	Is <e> *word* </e> used metaphorically?	*sentence*
f2	Is <e> *word* </e> used metaphorically in <s> *sentence* </s>?	*metaphor definition*
f3	Does <e> *word* </e> mean as if or like *word*?	*sentence*
f4	*word*	*sentence*

Table 1: Different formulations of questions for metaphor detection. For *metaphor definition* we use, “Metaphor is a figure of speech in which a word or phrase is applied to an object or action to which it is not literally applicable”, taken from Merriam-Webster.

struct two sets of questions and passages using the following pattern:

Question: Does <e> *word* </e> mean *definition*?
Passage: *sentence example*

The correct definition is used to form a pair of question and passage with a “Yes” answer, and a random definition is chosen from the rest of the glosses for *word* to form a question with a “No” answer. This construction requires WordNet entries with more than one definition. We split the dataset into a training set of 32K instances and evaluation set of 7.5K instances. Both splits are fully balanced in respect to positive and negative labels. Note that there is no overlap of *words* between the two splits.

4.2 VUA Metaphor Dataset

We train and evaluate on the widely used VUA corpus (Steen et al., 2010), with the splits provided in (Leong et al., 2020); see Table 2 for details. The metaphoric tokens that are annotated in this corpus are of four parts of speech: nouns, verbs, adjectives and adverbs. We use VUA in two different formats: the original, token classification format, and the yes–no question answering format, denoted VUA$_{qa}$.

5 Experiments

5.1 QA Models

We begin by training several QA models, each on a different dataset: (a) The **BoolQ** model is trained on the entire BoolQ data. (b) **WordNet** is trained on the entire WordNet. (c) **Mix** is trained on both the BoolQ and WordNet datasets.

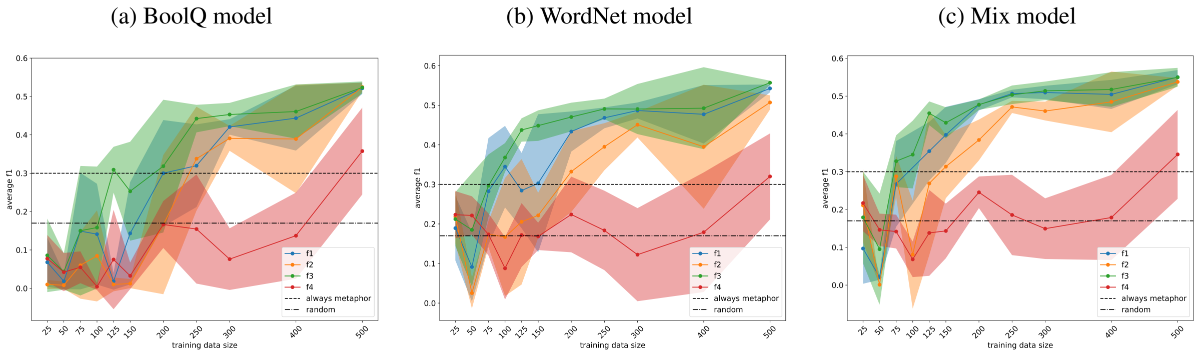
	Sentences	Tokens	Positive fraction
Train	12109	72611	18%
Test	4080	22198	17%

Table 2: Number of sentences, tokens and percentage of positive tokens in the VUA dataset.

The models are RoBERTa-base fine–tuned on two inputs – a question, followed by a passage (Devlin et al., 2019). We train for 10 epochs with batch size 32 and learning rate 1×10^{-5}. The number of training epochs is selected over the validation splits.

5.2 Metaphor Detection Models

We fine-tune each QA model on different subsets of VUA$_{qa}$, each subset of a different size, up to 500 sentences. Since each sentence contains multiple query tokens, for each sentence from VUA there are several training instance in VUA$_{qa}$, and thus 500 sentences annotated for metaphors on a token level transform into a few thousand training examples for all models that perform sequence classification for the single token in question. Each experiment is repeated four times with different random seeds, and we report the average F1 score and its standard deviation. In these experiments,


(a) BoolQ model

(b) WordNet model

(c) Mix model

Figure 2: Average F1 score over different training-set sizes, averaged over random seeds. Shaded areas indicate the standard deviation.

our aim is to analyze the competence of the underlying QA models in a zero- and few-shot scenario. We use a single set of parameters for all QA-based models. Specifically, a learning rate of 1×10^{-5}, batch size of 32, and 2 epochs. Following (Chen et al., 2020), we balance the weight at a ratio of 1:3 in favor of the positive label.

We use the following two baselines:

(a) A RoBERTa based sequence classifier that is fine-tuned on top of the RoBERTa pretrained model, similar to the baseline in (Choi et al., 2021). The input to this model is the concatenation of the sentence and the token in question, with the separation token in between. This baseline evaluates the contribution of the underlying question model. Note that this input is different than f4, since for f4, the token in question is the first input to the classifier. Since f4 is an input to a QA underlying model that accepts the question first, we maintain this order. However, for the baseline, since there is no QA model involved, we keep the recommended order for such classification tasks.

(b) DeepMet. For each dataset size, we fine tune four models with different random seeds and the results are averaged, similarly as for the QA-based models.

For the RoBERTa baseline, we tune hyper-parameters for each training data size with the technique suggested by (Zheng et al., 2022). Specifically, we experimented with batch size of 32, learning rate in $\{1\times10^{-5}, 3\times10^{-5}\}$ and number of training epochs in {2, 3}. DeepMet is evaluated with its default hyper-parameters. We also experimented with a RoBERTa token base classification, a baseline suggested in (Chen et al., 2020). While performing similarly to the sequence classifier when both were trained on the full data (Choi et al., 2021), for the few-shot scenario it is inferior to the sequence based classifier, and thus we omit it from the figures. We include the score of a classifier that randomly predicts “Metaphor” with the probability of the positive class over the entire dataset (18%), and the score of the classifier that always predicts “Metaphor”.

We begin with the evaluation of the different input patterns for our three models. Figure 2 shows the performance of the four patterns for each model. There is a clear advantage to all question-based patterns, with pattern f3 being the dominant one. Zero-shot is only relevant for our models, since the RoBERTa baseline is fine-tuned over a pretrained model and not a classification model. For all our models, the results in zero-shot mode are lower than the “always metaphor” baseline; thus, our architecture is not appropriate for this scenario.

Next, we assess the contribution of the underlying QA models. From Figure 2, we conclude that there is an overall advantage to the WordNet model over the BoolQ one across most patterns, and the Mix model is the best of all three.

In Figure 3, we compare our best model, namely, Mix with f3, with the best RoBERTa baseline and with DeepMet. Our model outperforms both baselines by a significant margin. In addition, we see a smaller standard deviation for our model, indicating that this architecture is more stable for small training datasets.

6 Conclusions and Future Work

QA-based models were shown to be effective for metaphor detection when training data is very limited. We analyzed the contribution of the question-like phrasing and the underlying QA model, and

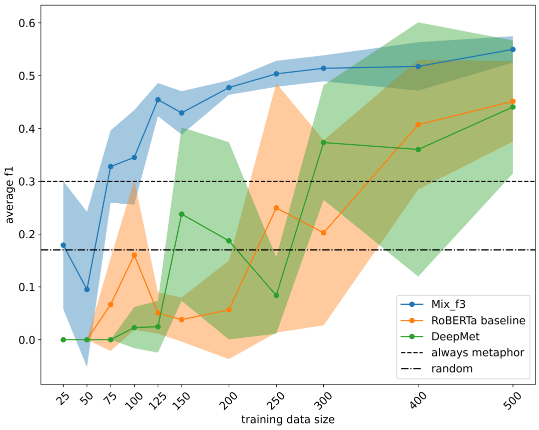


Figure 3: Mix model compared with the two baselines, RoBERTa and DeepMet.

report strong baselines for the few-shot scenario.

Another contribution is the use of WordNet. Transformer-based language models are pretrained on unlabeled data, thus many linguistic resources that were previously extensively used are less needed now. We have shown how the high-quality annotated data from WordNet can be utilized to train a QA model that can answer questions about semantics. We believe that similar techniques can generate high-quality datasets for training models for other NLP tasks.

As future work, we suggest exploring natural language inference models as underlying models for metaphor detection. Those models have been shown to be strong zero-shot models for various NLP tasks, so they can probably be of assistance in the metaphor domain. Another direction we aim to explore is the combination of our QA based technique with models such as DeepMet and MelBERT.

References

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pages 2037–2050, Dublin, Ireland. Association for Computational Linguistics.

Zaid Alyafeai, Maged Saeed AlShaibani, and Irfan Ahmad. 2020. A survey on transfer learning in natural language processing. *arXiv preprint arXiv:2007.04239*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference*, pages 9–15, Marseille, France. European Language Resource Association.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Xianyang Chen, Chee Wee (Ben) Leong, Michael Flor, and Beata Beigman Klebanov. 2020. Go figure! Multi-task transformer-based architecture for metaphor detection using idioms: ETS team in 2020 metaphor shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 235–243, Online. Association for Computational Linguistics.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936, Minneapolis, MN. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, MN. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Chee Wee Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 VUA and TOEFL metaphor detection shared task. In *Proceedings of* *the Second Workshop on Figurative Language Processing*, pages 18–29, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*, 1907.11692.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, volume Main, pages 255–269, Online. Association for Computational Linguistics.

Felix Schneider, Sven Sickert, Phillip Brandes, Sophie Marshall, and Joachim Denzler. 2022. Metaphor detection for low resource languages: From zero-shot to few-shot learning in Middle High German. In *LREC Workshop on Multiword Expression (LREC-WS)*, pages 75–80, Marseille, France. European Language Resources Association.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Shaked Refael Greenfeld, and Reut Tsarfaty. 2021. AlephBERT: A pre-trained language model to start off your Hebrew NLP application. *ArXiv*, 2104.04052.

Wei Song, Shuhui Zhou, Ruiji Fu, Ting Liu, and Lizhen Liu. 2021. Verb metaphor detection via contextual relation learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, volume 1: Long Papers, pages 4240–4251, Online. Association for Computational Linguistics.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*, volume 14 of *Converging Evidence in Language and Communication Research*. John Benjamins Publishing.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. DeepMet: A reading comprehension paradigm for token-level metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 30–39, Online. Association for Computational Linguistics.

Derek Tam, Rakesh R. Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. 2021. Improving and simplifying pattern exploiting training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4980–4991, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yanan Zheng, Jing Zhou, Yujie Qian, Ming Ding, Chonghua Liao, Li Jian, Ruslan Salakhutdinov, Jie Tang, Sebastian Ruder, and Zhilin Yang. 2022. FewNLU: Benchmarking state-of-the-art methods for few-shot natural language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pages 501–516, Dublin, Ireland. Association for Computational Linguistics.