Investigating BERT's Knowledge of Language: Five Analysis Methods with NPIs

Anonymous Preprint

Abstract

Though state-of-the-art sentence representation models can perform tasks requiring significant knowledge of grammar, it is an open question how best to evaluate their grammatical knowledge. We explore five experimental methods inspired by prior work evaluating pretrained sentence representation models. We use a single linguistic phenomenon, negative polarity item (NPI) licensing, as a case study for our experiments. NPIs like *any* are grammatical only if they appear in a *licensing environment* like negation (*Sue doesn't have any cats vs. *Sue has any cats*). This phenomenon is challenging because of the variety of NPI licensing environments that exist. We introduce an artificially generated dataset that manipulates key features of NPI licensing for the experiments. We find that BERT has significant knowledge of these features, but its success varies widely across different experimental methods. We conclude that a variety of methods is necessary to reveal all relevant aspects of a model's grammatical knowledge in a given domain.

1 Introduction

Recent sentence representation models have attained state-of-the-art results on language understanding tasks, but standard methodology for evaluating their knowledge of grammar has been slower to emerge. Recent work evaluating grammatical knowledge of sentence encoders like BERT (Devlin et al., 2018) has employed a variety of methods. For example, Shi et al. (2016), Ettinger et al. (2016), and Tenney et al. (2019) use probing tasks to target a model's knowledge of particular grammatical features. Marvin and Linzen (2018) and Wilcox et al. (2019) compare language models' probabilities for pairs of minimally different sentences differing in grammatical acceptability. Linzen et al. (2016), Warstadt et al. (2018), and Kann et al. (2019) use Boolean acceptability judgments inspired by methodologies in generative linguistics. However, we have not yet seen any substantial direct comparison between these methods, and it is not yet clear whether they tend to yield similar conclusions about what a given model knows.

We aim to better understand the trade-offs in task choice by comparing different methods inspired by previous work to evaluate sentence understanding models in a single empirical domain. We choose negative polarity item (*NPI*) licensing, an empirically rich phenomenon widely discussed in the theoretical linguistics literature, as our case study. NPIs are words or expressions that can only appear in environments that are, in some sense, *negative*. For example, *any* is an NPI because it is acceptable in negative sentences (1) but not positive sentences (2); negation thus serves as an NPI *licensor*. NPIs furthermore cannot be outside the syntactic *scope* of a licensor (3). Intuitively, a licensor's scope is the syntactic domain in which an NPI is licensed, and it varies from licensor to licensor. A sentence with an NPI present is only acceptable in cases where (i) there is a licensor—as in (1) but not (2)—and (ii) the NPI is within the scope of that licensor—as in (1) but not (3).

(1) Mary has**n't** eaten *any* cookies.

(2) *Mary has eaten *any* cookies.

(3) **Any* cookies have**n't* been eaten.

We compare five experimental methods to test BERT's knowledge of NPI licensing. We consider: (i) a *Boolean acceptability classification* task to test BERT's knowledge of sentences in isolation, (ii) an *absolute minimal pair* task evaluating whether the absolute Boolean outputs of acceptability classifiers distinguish between minimally different pairs of sentences, (iii) a *gradient*

minimal pair task evaluating whether the gradient outputs of acceptability classifiers distinguish between minimal pairs, (iv) a *cloze test* evaluating the grammatical preferences of BERT's masked language modeling head, and (v) a *probing task* evaluating BERT's representations for knowledge of specific grammatical features relevant to NPI licensing.

We find that BERT knows about NPI licensing environments. However, our five methods give meaningfully different results. In particular, the gradient minimal pair experiment leads us to believe that BERT has systematic knowledge about all NPI licensing environments and relevant grammatical features, while the absolute minimal pair and probing experiments show that BERT's knowledge is in fact not equal across these domains. We conclude that no single method is able to accurately depict all relevant aspects of a model's grammatical knowledge; comparing both gradient and absolute measures of performance of trained models gives a more complete picture. We recommend that future studies would benefit from using multiple converging methods to evaluate model performance.

2 Related Work

Evaluating Sentence Encoders The success of sentence encoders and broader neural network methods in NLP has prompted significant interest in understanding the linguistic knowledge encapsulated in these models.

A section of related work focuses on Boolean classification tasks to evaluate the grammatical knowledge encoded in these models. Linzen et al. (2016) uses acceptability classification of sentences with manipulated verbal inflection to investigate whether LSTMs can identify subject-verb agreement violations, and therefore a (potentially long distance) syntactic dependency. Warstadt et al. (2018) uses sentence acceptability on a corpus of judgments as a task for evaluating grammatical knowledge. Kann et al. (2019) introduces methods for testing whether word and sentence encoders represent information about verbal argument structure.

Marvin and Linzen (2018) and Wilcox et al. (2019) employ minimally different sentences in terms of linguistic acceptability to judge whether the encoder is sensitive to this ungrammatically.

Another branch of work uses probing classifiers to reveal how much information a sentence embedding encodes about syntactic and surface features such as tense and voice (Shi et al., 2016), sentence length and word content (Adi et al., 2016), or syntactic depth and morphological number (Conneau et al., 2018). Giulianelli et al. (2018) use diagnostic classifiers to track the propagation of information in RNN-based language models. Ettinger et al. (2018) and Dasgupta et al. (2018) use automatic data generation to evaluate compositional reasoning.

To study contextualized sentence encoders Devlin et al. (2018); Peters et al. (2018); Radford et al. (2018), Tenney et al. (2019) introduce subsentence level edge probing tasks derived from NLP tasks, providing evidence that these encoders trained on language modeling and translation encode more syntax than semantics.

Negative Polarity Items In the theoretical literature on NPIs, proposals have been made to unify the properties of the diverse NPI licensing environments. For example, a popular view states that NPIs are licensed if they occur in downward entailing environments (Fauconnier, 1975; Ladusaw, 1979), i.e. an environment that licences inferences from sets to subsets.[1]

Within computational linguistics, Marvin and Linzen (2018) find that LSTMs do not systematically assign a higher probability to grammatical sentences like (1) than minimally different ungrammatical sentences like (2). Wilcox et al. (2019) use NPIs, along with filler-gap dependencies, as instances of non-local grammatical dependencies, to probe the effect of supervision with hierarchical structure. They find that structurally-supervised models outperform state-of-the-art sequential LSTM models, showing the importance of structure in learning non-local dependencies like NPI licensing.

Acceptability Judgments The ability of neural networks to make Boolean acceptability judgments was previously studied using the Corpus of Linguistic Acceptability (CoLA; Warstadt et al., 2018). CoLA consists of over 10k syntactically diverse example sentences from the linguistics literature with expert acceptability labels. As is conventional in theoretical linguistics, sentences

[1]Other prominent theories of NPI licensing are based on notions of non-veridicality (Giannakidou, 1994, 1998; Zwarts, 1998), domain widening and emphasis (Kadmon and Landman, 1993; Krifka, 1995; Chierchia, 2013), a.o.

are taken to be *acceptable* if native speakers judge them to be possible sentences in their language. Such sentences are widely used in linguistics publications to illustrate phenomena of interest. The examples in CoLA are gathered from diverse sources and represent a wide array of syntactic, semantic, and morphological phenomena. As measured by the GLUE benchmark (Wang et al., 2018), acceptability classifiers trained on top of BERT and related models reach near-human performance on CoLA.

3 Methods

We experiment with five approaches to the evaluation of grammatical knowledge of sentence representation models like BERT using our generated NPI acceptability judgment dataset (§4). Each data sample in the dataset contains a sentence, a Boolean label which indicates whether the sentence is grammatically acceptable or not, and three Boolean meta-data variables (licensor, NPI, and scope; Table 2).

Boolean Acceptability We test the model's ability to judge the grammatical acceptability of the sentences in the NPI dataset. Following standards in linguistics,sentences for this task are assumed to be either totally acceptable or totally unacceptable. We fine-tune the sentence representation models to perform these Boolean judgments. For BERT-based sentence representation models, we add a classifier on top of `[CLS]` embedding of the last layer. For BoW, we use a max pooling layer followed by an MLP classifier. The performance of the models is measured as Matthews Correlation Coefficient (MCC; Matthews, 1975)[2] between the predicted label and the gold label.

Absolute Minimal Pair We conduct a minimal pair experiment to analyze Boolean acceptability classifiers on minimally different sentences. Two sentences form a minimal pair if they differ in only one NPI-related Boolean meta-data variable within a paradigm, but have different acceptability. We evaluate the models trained on acceptability judgments with the minimal pairs. In *absolute minimal pair* evaluation, the models need to correctly classify both sentences in the pair to be counted as correct.

[2]MCC gives the correlation of two Boolean distributions between -1 and 1. On average, a score of 0 will be given to any two unrelated distributions, regardless of class imbalance.

Gradient Minimal Pair The *gradient minimal pair* evaluation is a more lenient version of *absolute minimal pair* evaluation: Here, we count a pair as correctly classified as long as the model predicts that the probability of the grammatically acceptable sentence is higher that that of the ungrammatical sentence.

Cloze Test In the cloze test, a standard sentence-completion task, we use the masked language modeling (MLM) component in BERT Devlin et al. (2018) and evaluate whether it assigns a higher probability to the grammatical sentence in a minimal pair, following (Linzen et al., 2016). An MLM predicts the probability of a single masked token based on the rest of the sentence. The minimal pairs tested are a subset of those in the absolute and gradient minimal pair experiments, where both sentences must be equal in length and differ in only one token. This differing token is replaced with `[MASK]`, and the minimal pair is taken to be classified correctly if the MLM assigns a higher probability to the token from the acceptable sentence. In contrast with the other minimal pair experiments, this experiment is entirely unsupervised, using BERT's native MLM functionality.

Feature Probing We use probing classifiers as a more fine-grained approach to the identification of grammatical variables. We freeze the sentence encoders both with and without fine-tuning from the acceptability judgment experiments and train lightweight classifiers on top of them to predict meta-data labels (licensor, NPI, and scope). Crucially, each individual meta-data label by itself does not decide acceptability (i.e., these probing experiments test a different but related set of knowledge from acceptability experiments).

4 Data

In order to probe BERT's performance on sentences involving NPIs, we generate a set of sentences and acceptability labels for the experiments in this paper. We use generated data so that we can assess minimal pairs, and so that there are sufficient unacceptable sentences.

Data generation We create a controlled set of 136,000 English sentences using an automated sentence generation procedure, inspired in large part by previous work by Ettinger et al. (2016, 2018), Marvin and Linzen (2018), Dasgupta et al. (2018), and Kann et al. (2019). The set contains

Environment	Abbrev.	Example
Adverbs	ADV	The guests who **rarely** love *any* actors had left libraries.
Conditionals	COND	**If** the pedestrian passes *any* schools, the senator will talk to the adults.
Determiner negation	D-NEG	Just as the waitress said, **no** customers thought that *any* dancers bought the dish.
Sentential negation	S-NEG	These drivers have **not** thought that *any* customers have lied.
Only	ONLY	From what the cashier heard, **only** the children have known *any* dancers.
Quantifiers	QNT	**Every** actress who was talking about *any* high schools criticizes the children.
Questions	QUES	The boys wonder **whether** the doctors went to *any* art galleries.
Simple questions	SMP-Q	**Has** the guy worked with *any* teenagers**?**
Superlatives	SUP	The teenagers approach the **nicest** actress that *any* customers had criticized.

Table 1: Examples of each of the NPI licensing environments generated. The licensor is in bold, and the NPI (here *any*) is in italics. All examples show cases where the NPI is present (NPI=1), the licensor is present (Licensor=1), and the NPI is in the scope of the licensor (Scope=1); all are acceptable to native speakers of English.

Licensor	NPI	Scope	Sentence
1	1	1	Those boys wonder **whether** [the doctors went to *any* art galleries.]
1	1	0	**Any* boys wonder **whether** [the doctors went to art galleries.]
1	0	1	Those boys wonder **whether** [the doctors went to *the* art galleries.]
1	0	0	*The* boys wonder **whether** [the doctors went to art galleries.]
0	1	1	*Those boys say **that** [the doctors went to *any* art galleries.]
0	1	0	**Any* boys say **that** [the doctors went to art galleries.]
0	0	1	Those boys say **that** [the doctors went to *the* art galleries.]
0	0	0	*The* boys say **that** [the doctors went to art galleries.]

Table 2: Example 2×2×2 paradigm using the Questions environment. The licensor (*whether*) or licensor replacement (*that*) is in bold. The NPI (*any*) or NPI replacement (*the*) is in italics. When licensor=1, the licensor is present rather than its replacement word. When NPI=1, the NPI is present rather than its replacement. The scope of the licensor/licensor replacement is shown in square brackets (brackets, italicization, and boldface are not present in the actual data). When scope=1, the NPI/NPI replacement is within the scope of the licensor/licensor replacement. Unacceptable sentences are marked with *. The five minimal pairs are connected by arrows that point from the unacceptable to the acceptable sentence.

nine NPI licensing environments (Table 1), and two NPIs (*any*, *ever*). All but one licensor-NPI pair follows a 2×2×2 paradigm, which manipulates three variables: licensor presence, NPI presence, and the occurrence of an NPI within a licensor's scope. Each 2×2×2 paradigm forms 5 minimal pairs. Table 2 shows an example paradigm.

Licensor presence indicates whether an NPI licensor is in the sentence. When the licensor is not present, it is replaced by a word that does not license NPIs but has a similar structural distribution. Similarly, NPI presence indicates whether an NPI is in the sentence or if it is replaced by a non-NPI that has a similar structural distribution. Scope indicates whether the NPI/NPI replacement is within the scope of the licensor/licensor replacement. The scope manipulation indicates whether an NPI occurs within the syntactic scope of its licensor. As illustrated earlier in (3), a sentence containing an NPI is only acceptable when the NPI falls within the scope of the licensor.

The exception to the 2×2×2 paradigm is the Simple Questions licensing condition, with a reduced 2×2 paradigm. It lacks a scope manipulation because the question takes scope over the entire clause, and in Simple Questions the clause is the whole sentence. The paradigm for Simple Questions is given in Table 3 in the Appendix, it forms only 2 minimal pairs.

To generate the sentences, we create sentence templates for each paradigm. Templates follow the general structure illustrated in example (4), in which the part-of-speech (auxiliary verb, determiner, noun, verb), as well as the instance number is specified. For example, N2 is the second instance of a noun in the template. We use these labels here for illustrative purposes; in reality, the templates also include more fine-grained specifications, such as verb tense and noun number.

(4)
Aux1 D1 N1 V1 any N2 ?
Has the guy seen any waitresses ?

Given the specifications encoded in the sentence templates, words were sampled from a vocabulary

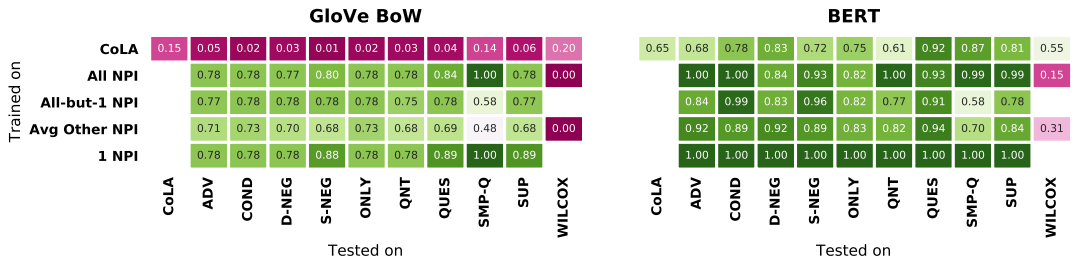

Figure 1: Results from the acceptability judgment experiment in MCC. The columns indicate evaluation tests, and the rows fine-tuning settings.

of over 1000 lexical items annotated with 30 syntactic, morphological, and semantic features. The annotated features allow us to encode selectional requirements of lexical items, e.g. what types of nouns a verb can combine with.[3] This avoids blatantly implausible sentences.

For each environment, the training set contains 10K sentences, and the dev and test sets contain 1K sentences each. Sentences from the same paradigm are always in the same set.

In addition to our data set, we also test BERT on a set of 104 handcrafted sentences from the NPI sub-experiment in Wilcox et al. (2019), who use a paradigm that partially overlaps with ours, but has an additional condition where the NPI linearly follows its licensor while not being in the scope of the licensor. This is included as an additional test set for evaluating acceptability classifiers in (6).

Data validation We use Amazon Mechanical Turk (MTurk) to validate a subset of our sentences to assure that the generated sentences represent a real contrast in acceptability. We randomly sample five-hundred sentences from the dataset, sampling approximately equally from each environment, NPI and paradigm. Each sentence is rated by 20 different participants on a Likert scale of 1-6, with 1 being "the sentence is not possible in English" and 6 being "the sentence is possible in English". A Wilcoxon signed-rank test (Wilcoxon, 1945) shows that within each environment and for each NPI, the acceptable sentences are more often rated as acceptable by our MTurk validators than the unacceptable sentences (all p-values < 0.001). This contrast holds considering both the raw Likert-scale responses and the responses transformed to a Boolean judgment. Table 4 in the Appendix shows the participants' scores transformed into a Boolean judgment of 0 (unacceptable, score ≤ 3) or 1 (acceptable, score ≥ 4) and presented as the percentage of 'acceptable' ratings assigned to the sentences in a given condition.

5 Experimental Settings

We conduct our experiments with the jiant 0.9 (Wang et al., 2019) multitask learning and transfer learning toolkit, the AllenNLP platform (Gardner et al., 2018), and the BERT implementation from HuggingFace.[4]

Models We study the following sentence understanding models: (i) *GloVe BoW*: a bag-of-words baseline obtained by max-pooling of 840B tokens 300-dimensional GloVe word embeddings (Pennington et al., 2014) and (ii) *BERT* (Devlin et al., 2018): we use the cased version of BERT-large model, which works the best for our tasks in pilot experiments. In addition, since recent work (Liu et al., 2019; Stickland and Murray, 2019) has shown that intermediate training on related tasks can meaningfully impact BERT's performance on downstream tasks, we also explore two additional BERT-based models—(iii) *BERT→MNLI*: BERT fine-tuned on the Multi-Genre Natural Language Inference corpus (Williams et al., 2018), motivated both by prior work on pretraining sentence encoders on MNLI (Conneau et al., 2017) as well as work showing significant improvements to BERT on downstream semantic tasks (Phang et al., 2018; Bowman et al., 2018) (iv) *BERT→CCG*: BERT fine-tuned on Combinatory Categorial Grammar Bank corpus (Hockenmaier and Steedman, 2007), motivated by Wilcox et al.'s (2019) finding that structural supervision may improve a LSTM-based sentence encoders knowledge on non-local syntactic dependencies.

[3] All data and code used in sentence generation is available on GitHub, and will be linked to upon acceptance.

[4] https://github.com/huggingface/pytorch-pretrained-BERT

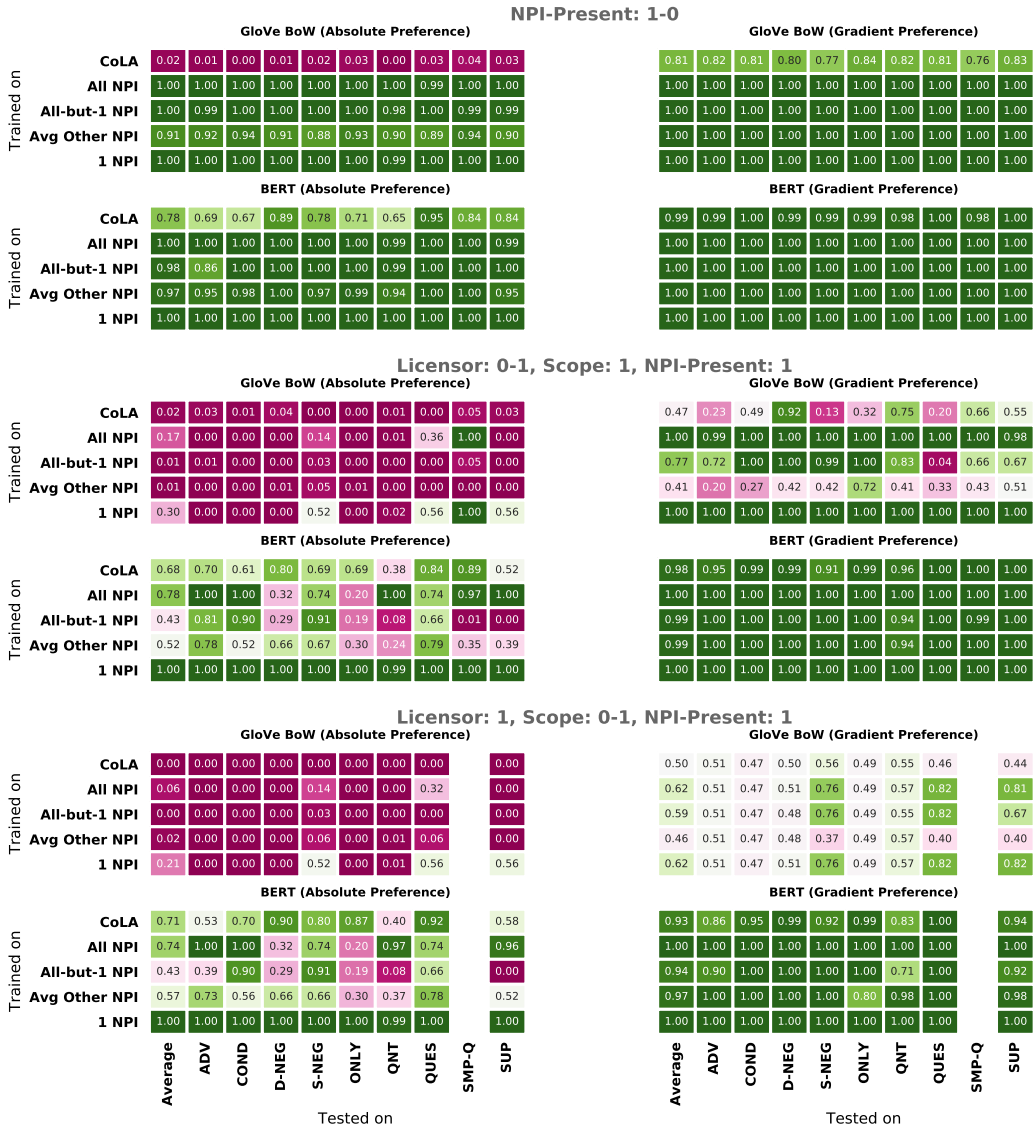

Figure 2: Results from the minimal pair test. The top section shows the average accuracy for NPI detection, the middle section shows average accuracy for licensor detection, and the bottom shows average accuracy of minimal pair contrasts that differ scope. Within each section, we show performance of GloVe BoW and BERT models under both absolute preference and gradient preference evaluation methods. The rows represent the training-evaluation configuration, while the columns represent different licensing environments.

Training-Evaluation Configurations We are interested in whether sentence representation models learn NPI licensing as a unified property. Can the models generalize from trained environments to previously unseen environments? To answer these questions, for each NPI environment, we extensively test the performance of the models in the following configurations: (i) *CoLA*: training on CoLA, evaluating on the environment. (ii) *1 NPI*: training and evaluating on the same NPI environment. (iii) *Avg Other NPI*: training independently on every NPI environment except one, averaged over the evaluation results on that environment. (iv) *All-but-1 NPI*: training on all environments except for one environment, evaluating on that environment. (v) *All NPI*: training on all environments, evaluating on the environment.

6 Results

Acceptability Judgments The results in Fig. 1 show that BERT outperforms the BoW baseline on all test data with all fine-tuning settings. Within each BERT variants, MCC reaches 1.0 on all test data in the *1 NPI* setting. When the *All-but-1 NPI* training-evaluation configuration is used, the performance on all NPI environments for BERT

drops. While the MCC value on environments like conditionals and sentential negation remains above 0.9, on the simple question environment it drops to 0.58. Compared with NPI data fine-tuning, CoLA fine-tuning results in BERT's lower performance on most of the NPI environments but better performance on data from Wilcox et al. (2019).

In comparing the three BERT variants (see full results in Figure 5 in the Appendix), the *Avg Other NPI* shows that on 7 out of 9 NPI environments, plain BERT outperforms BERT→MNLI and BERT→CCG. Even in the remaining two environments, plain BERT yields about as good performance as BERT→MNLI and BERT→CCG, indicating that MNLI and CCG fine-tuning brings no obvious gain to acceptability judgments.

Absolute and Gradient Minimal Pairs The results (Fig. 2) show that models' performance hinges on how minimal pairs differ. When tested on minimal pairs differing by the presence of an NPI, BoW and plain BERT obtain (nearly) perfect accuracy on both absolute and gradient measures across all settings. For minimal pairs differing by licensor and scope, BERT again achieves near perfect performance on the gradient measure, while BoW does not. On the absolute measure, both BERT and BoW perform worse. Overall, it shows that absolute judgment is more challenging when targeting licensor, which involves a larger pool of lexical items and syntactic configurations than NPIs, and scope, which requires nontrivial syntactic knowledge about NPI licensing.

As in the acceptability experiment, we find that intermediate fine-tuning on MNLI and CCG does not improve performance (see full results in Figures 6-8 in Appendix).

Cloze Test The results (Fig. 3) show that even without supervision on NPI data, the BERT MLM can distinguish between acceptable and unacceptable sentences in the NPI domain. Performance is highly dependent on the NPI environment and type of minimal pair. Accuracy for NPI-detection falls between 0.76 and 0.93 for all environments. Accuracy for licensor-detection is much more variable, with the BERT MLM achieving especially high performance in conditional, sentential negation, and *only* environments; and low performance in quantifier and superlative environments.

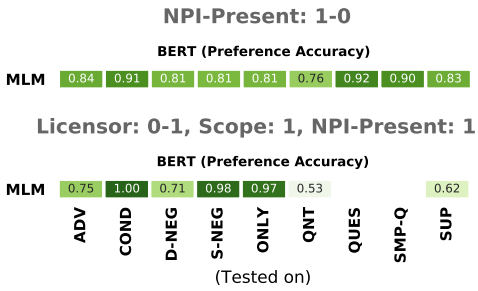

Figure 3: Results of BERT MLM performance in the cloze test. The top section shows the average accuracy for NPI detection; the bottom section show the accuracy for licensor detection. The columns represent different licensing environments

Feature Probing Results (Fig. 4) show that plain BERT outperforms the BoW baseline in detecting scope. As expected, BoW is nearly perfect in detecting presence of NPI and licensor, as these tasks do not require knowledge of syntax or word order. Consistent with results from previous experiments, licensor detection is slightly more challenging for models fine-tuned with CoLA or NPI data. However, the overall lower performances in scope detection compared with licensor detection is not found in the minimal-pair experiments.

CoLA fine-tuning improves the performance for BERT, especially for NPI presence. Fine-tuning on NPI data improves scope detection. Inspection of environment-specific results shows that models struggle when the superlative, quantifiers, and adverb environments are the held-out test sets in the All-but-1 NPI fine-tuning setting.

Different from other experiments, BERT and BERT→MNLI have comparable performance across many settings and tasks, beating BERT→CCG especially in scope detection (see full results in Figure 9 in the Appendix).

7 Discussion

We find that BERT systematically represents all features relevant to NPI licensing across most environments according to certain evaluation methods. However, these results vary widely across the different methods we compare. In particular, BERT performs nearly perfectly on the gradient minimal pairs task (at ceiling) across all of minimal pair configurations and nearly all licensing environments. Based on this method alone, we might conclude that BERT's knowledge of this domain is near perfect. However, the other methods show a

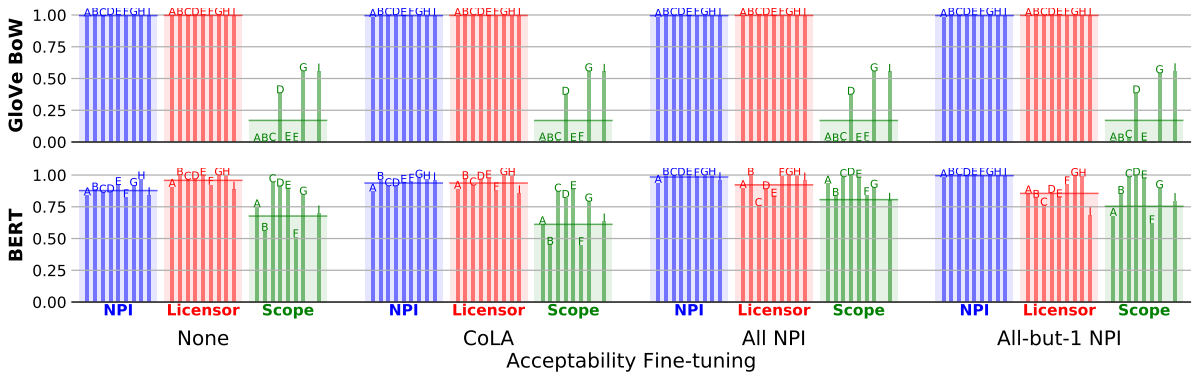

Figure 4: Results of probing classification on NPI presence, licensor presence, and scope detection, shown in MCC. Letters on top of bars refer to NPI environments: A=ADV, B=COND, C=D-NEG, D=S-NEG, E=ONLY, F=QNT, G=QUES, H=SMP-Q, I=SUP.

more nuanced picture.

BERT's knowledge of which expressions are NPIs and NPI licensors is generally stronger than its knowledge of the licensors' scope. This is especially apparent from the probing results (Fig. 4). BERT without acceptability fine-tuning performs close to ceiling on the licensor-detection probing task, but is inconsistent at scope-detection. Tellingly, the BoW baseline is also able to perform at ceiling on the and licensor-detection probing task. For BoW to succeed at this task, the GloVe embeddings for NPI-licensors must share some common property, most likely the fact that licensors co-occur with NPIs. It is possible that BERT is able to succeed using a similar strategy. By contrast, identifying whether an NPI is in the scope of a licensor requires at the very least word order information and not just co-occurrences.

The contrast in BERT's performance on the gradient and absolute tasks tells us that these evaluations reveal different aspects of BERT's knowledge. The gradient task is strictly easier than the absolute task. On the one hand, BERT's high performance on the gradient task reveals the presence of systematic knowledge in the NPI domain. On the other hand, due to ceiling effects, the gradient task fails to reveal actual differences between environments that we clearly observe based on absolute, cloze, and probing tasks.

While BERT has systematic knowledge of acceptability contrasts, this knowledge varies across environments and is not categorical. Current linguistic theory models human knowledge of natural language as categorical: In that sense BERT fails at attaining human performance. However, it is unclear whether humans themselves achieve categorical performance. Results from an MTurk study on human acceptability of our generated dataset show non-categorical agreement with the judgments in our dataset.

Supplementing BERT with additional pretraining on CCG and MNLI does not improve performance, and even lowers performance in some cases. While results from Phang et al. (2018) lead us to hypothesize that intermediate pretraining might help, this is not what we observe on our data. This result is in direct contrast with the results from Wilcox et al. (2019), who find that syntactic pretraining does improve performance in the NPI domain. This difference in findings is likely due to differences in models and training procedure, as their model is an RNN jointly trained on language modeling and parsing over the much smaller Penn Treebank (Marcus et al., 1993).

Future studies would benefit from employing a variety of different methodologies for assessing model performance withing a specified domain. In particular, a result showing generally good performance for a model should be regarded as possibly hiding actual differences in performance that a different task would reveal. Similarly, generally poor performance for a model does not necessarily mean that the model does not have systematic knowledge in a given domain; it may be that an easier task would reveal systematicity.

8 Conclusion

We have shown that within a well-defined domain of English grammar, evaluation of sentence encoders using different tasks will reveal different aspects of the encoder's knowledge in that domain.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.

Samuel R. Bowman, Ellie Pavlick, Edouard Grave, Benjamin Van Durme, Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R. Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, Shuning Jin, and Berlin Chen. 2018. Looking for elmo's friends: Sentence-level pretraining beyond language modeling. *CoRR*, abs/1812.10860.

Gennaro Chierchia. 2013. *Logic in grammar: Polarity, free choice, and intervention*. Oxford University Press.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *CoRR*, abs/1705.02364.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J Gershman, and Noah D Goodman. 2018. Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint 1810.04805*.

Allison Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the First Workshop on Evaluating Vector Space Representations for NLP*, pages 134–139. Association for Computational Linguistics.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801. Association for Computational Linguistics.

Gilles Fauconnier. 1975. Polarity and the scale principle. *Proceedings of the Chicago Linguistics Society*, 11:188–199.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6.

Anastasia Giannakidou. 1994. The semantic licensing of NPIs and the Modern Greek subjunctive. *Language and Cognition*, 4:55–68.

Anastasia Giannakidou. 1998. *Polarity sensitivity as (non) veridical dependency*, volume 23. John Benjamins Publishing.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem H. Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. *CoRR*, abs/1808.08079.

Julia Hockenmaier and Mark Steedman. 2007. CCGbank: A corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics*, 33(3):355–396.

Nirit Kadmon and Fred Landman. 1993. Any. *Linguistics and Philosophy*, 16(4):353–422.

Katharina Kann, Alex Warstadt, and Adina Williams. 2019. Verb argument structure alternations in word and sentence embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL)*, pages 287–297.

Manfred Krifka. 1995. The semantics and pragmatics of polarity items. *Linguistic Analysis*, 25(3-4):209–257.

William Ladusaw. 1979. Negative polarity items as inherent scope relations. *Unpublished Ph.D. Dissertation, University of Texas at Austin*.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *CoRR*, abs/1901.11504.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL*.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence encoders on STILTs: Supplementary training on intermediate labeled-data tasks. In *arXiv preprint*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Unpublished manuscript accessible via the OpenAI Blog.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural mt learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534.

Asa Cooper Stickland and Iain Murray. 2019. BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning. *CoRR*, abs/1902.02671.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Alex Wang, Ian F. Tenney, Yada Pruksachatkun, Katherin Yu, Jan Hula, Patrick Xia, Raghu Pappagari, Shuning Jin, R. Thomas McCoy, Roma Patel, Yinghui Huang, Jason Phang, Edouard Grave, Najoung Kim, Phu Mon Htut, Thibault F'evry, Berlin Chen, Nikita Nangia, Haokun Liu, Anhad Mohananey, Shikha Bordia, Ellie Pavlick, and Samuel R. Bowman. 2019. jiant 0.9: A software toolkit for research on general-purpose text understanding models. `http://jiant.info/`.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2018. Neural network acceptability judgments. *arXiv preprint 1805.12471*.

Ethan Wilcox, Peng Qian, Richard Futrell, Miguel Ballesteros, and Roger Levy. 2019. Structural supervision improves learning of non-local grammatical dependencies. In *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Frans Zwarts. 1998. Three types of polarity. In *Plurality and Quantification*, pages 177–238. Springer.

Appendix

Lic.	NPI	Sentence
1	1	**Has** the guy worked with *any* teenagers?
1	0	**Has** the guy worked with *the* teenagers?
0	1	*The guy **has** worked with *any* teenagers.
0	0	The guy **has** worked with *the* teenagers.

Table 3: Reduced paradigm for Simple questions. "Lic." is abbreviated from "Licensor". The licensor and licensor replacement are shown in bold (*has* in both cases). The NPI (*any*) and NPI replacement (*the*) are shown in italics. There is no scope manipulation because it is not possible to place an NPI or NPI replacement outside of the scope of an interrogative or declarative phrase. The 2 minimal pairs are shown by arrows, pointing from unacceptable to acceptable sentence.

Environment	Label	% accept	Diff
Adverb	*	8.33	61.67
	✓	70.00	
Conditionals	*	37.50	50.00
	✓	87.50	
Determiner negation	*	11.11	78.89
	✓	90.00	
Embedded questions	*	8.33	89.17
	✓	97.50	
Only	*	5.56	84.44
	✓	90.00	
Sentential negation	*	27.78	52.22
	✓	80.00	
Simple questions	*	33.33	62.97
	✓	96.30	
Superlatives	*	8.33	66.67
	✓	75.00	
Quantifiers	*	4.17	50.83
	✓	55.00	

Table 4: Results from MTurk validation. 'Environment' is the name of the licensing environment and 'label' is whether the sentence was intended as acceptable (✓) or unacceptable (*). The results of the validation ratings is in '% accept' and represents the majority vote for each sentence as acceptable/unacceptable and then averaged to give the percentage of times a sentence in a given condition was rated as acceptable by the MTurk raters. 'Diff' is calculated from the % of acceptable sentences rated acceptable minus the % of unacceptable sentences rated acceptable (100 is a perfect score, 0 means there is no difference).

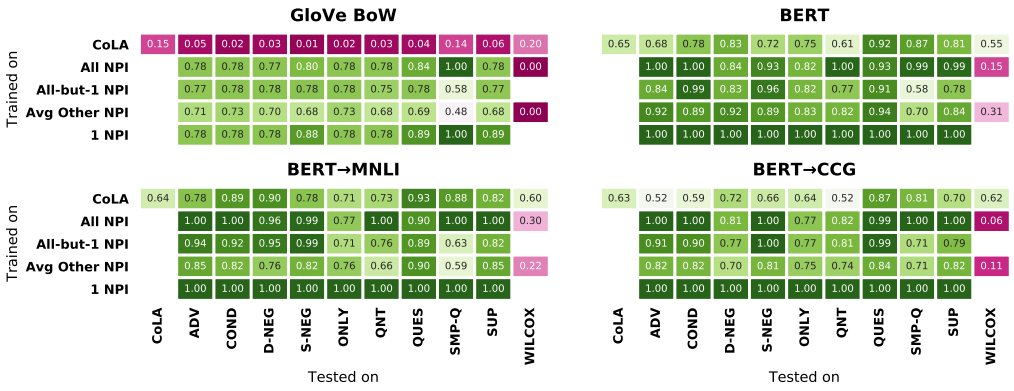

Figure 5: Results from the acceptability judgment experiment in MCC. The columns indicate evaluation tests, and the rows fine-tuning settings.

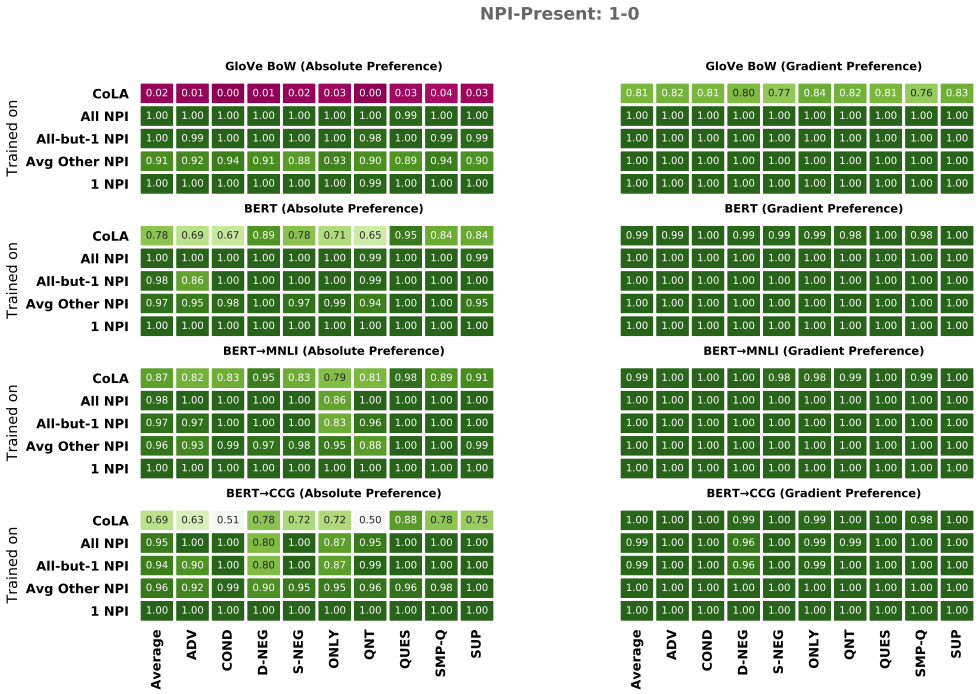

Figure 6: Results from minimal pair test for the NPI-presence contrast. The smaller diagrams of each sector show performance of BoW and BERT variants under two different minimal pair evaluation methods. The rows represent training-evaluation configuration, while the columns represent different licensing environments.

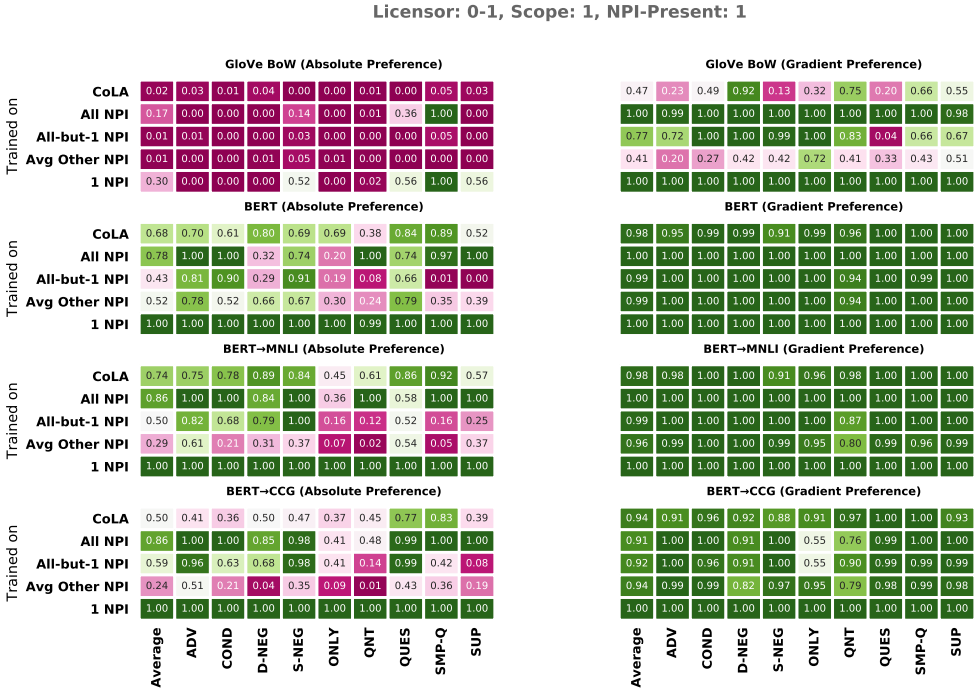

Figure 7: Results from minimal pair test for the licensor-presence contrast. The smaller diagrams of each sector show performance of BoW and BERT variants under two different minimal pair evaluation methods. The rows represent training-evaluation configuration, while the columns represent different licensing environments.

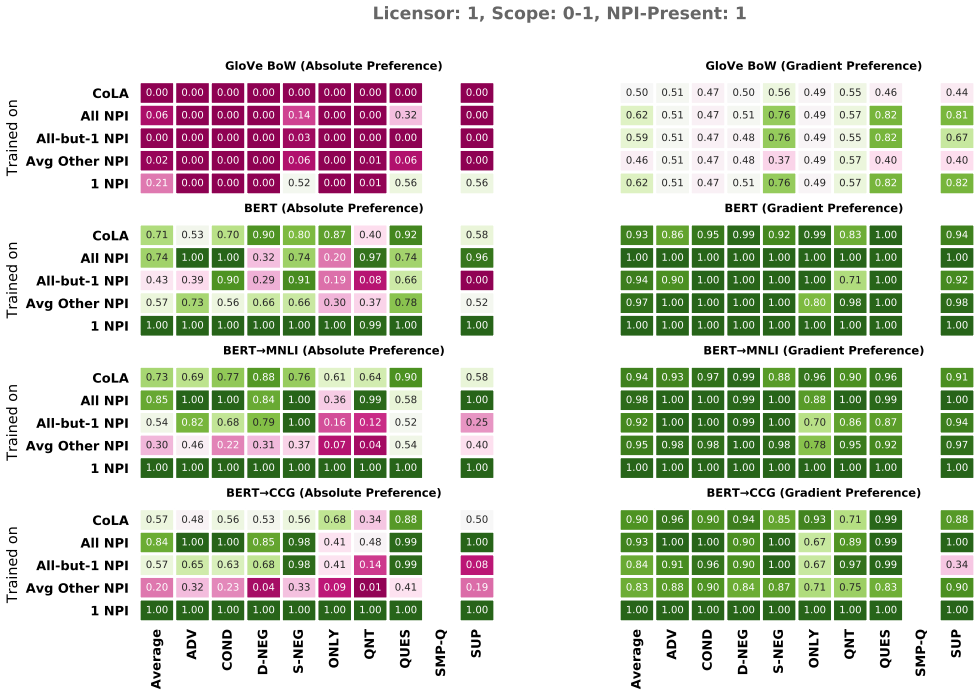

Figure 8: Results from minimal pair test for the scope contrast. The smaller diagrams of each sector show performance of BoW and BERT variants under two different minimal pair evaluation methods. The rows represent training-evaluation configuration, while the columns represent different licensing environments.

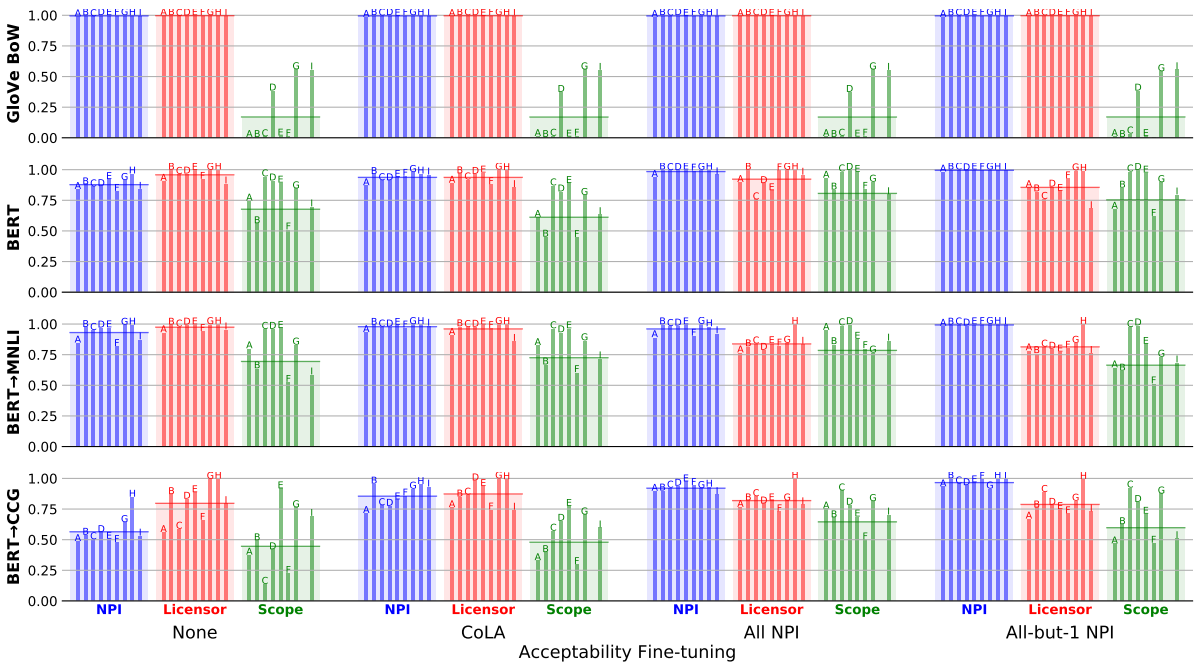

Figure 9: Results of probing classification on NPI presence, licensor presence, and scope detection, shown in MCC. Letters on top of bars refer to NPI environments: A=ADV, B=COND, C=D-NEG, D=S-NEG, E=ONLY, F=QNT, G=QUES, H=SMP-Q, I=SUP.