Probing What Different NLP Tasks Teach Machines about Function Word Comprehension

Najoung Kim[†], **Roma Patel**[*], **Adam Poliak**[△], **Alex Wang**[∂], **Patrick Xia**[△], **Tom McCoy**[†], **Ian Tenney**[‡], **Alexis Ross**[◇], **Tal Linzen**[†], **Benjamin Van Durme**[†,△], **Sam Bowman**[∂], **Ellie Pavlick**[*]

[†]Dept. of Cognitive Science, Johns Hopkins University
[△]Dept. of Computer Science, Johns Hopkins University
[*]Dept. of Computer Science, Brown University
[∂]Dept. of Computer Science, New York University
[‡]Google AI Language
[◇]Dept. of Computer Science, Harvard University

Abstract

We introduce a set of nine challenge tasks that test for the understanding of function words. These tasks are created by structurally mutating sentences from existing datasets to target the comprehension of specific types of function words (e.g., prepositions, *wh*-words). Using these probing tasks, we explore the effects of various pretraining objectives for sentence encoders (e.g., language modeling, CCG supertagging and natural language inference (NLI)) on the learned representations. Our results show that pretraining on CCG—our most syntactic objective—performs the best on average across our probing tasks, suggesting that syntactic knowledge helps function word comprehension. Language modeling also shows strong performance, supporting its widespread use for pretraining state-of-the-art NLP models. Overall, no pretraining objective dominates across the board, and our function word probing tasks highlight several intuitive differences between pretraining objectives, e.g. that NLI helps the comprehension of negation.

1 Introduction

Many recent advances in NLP have been driven by new approaches to representation learning—i.e. the design of models whose primary aim is to yield representations of words or sentences that useful for a range of downstream applications (Bowman et al., 2017). Approaches to representation learning typically differ in either the architecture of the model used to learn the representations, the objective used to train that network, or both. Varying these factors can significantly impact performance on a broad range of NLP tasks (McCann et al., 2017; Peters et al., 2018; Devlin et al., 2019).

This paper investigates the role of pretraining objectives of sentence encoders, with respect to their capacity to understand function words (e.g., prepositions, conjunctions). Although the importance of finding an effective pretraining objective is well acknowledged, relatively few studies offer a controlled comparison of diverse pretraining objectives, holding model architecture constant.

We ask whether the linguistic properties implicitly captured by pretraining objectives measurably affect the types of linguistic information encoded in the learned representations. To this end, we explore whether qualitatively different objectives lead to demonstrably different sentence representations. We focus our analysis on function words because they play a key role in compositional meaning—e.g., introducing and identifying discourse referents or representing relationships between entities or ideas—and are not yet considered to be well-modeled by distributional semantics (Bernardi et al., 2015). Our results suggest that different pretraining objectives give rise to differences in function word comprehension; for instance, we see that natural language inference helps understanding negation, and grounded language helps understanding spatial descriptors. However, overall, we find that the observed differences are not always straightforwardly interpretable, and further investigation is needed to determine what specific aspects of pretraining tasks, yield good representations of function words.

The analyses we present contribute new results in an ongoing line of research aimed at providing a finer-grained understanding of what neural networks capture about linguistic structure (Conneau et al., 2018; Poliak et al., 2018; Linzen et al., 2018; Tenney et al., 2019, *i.a.*). Our contributions are:

- We provide an in-depth exploration into how different pretraining objectives for sentence encoders affect the information encoded by the output representations. We isolate the effects of different pretraining objectives by

	Acceptability			
wh-word	**why** are you so chippy about posh people?			✓
	. . .a Mr. Nice Guy like Melcher, **what** is now 46			X
Def.	. . . **the** case is remarkable for **the** cooperation . . .			✓
	. . . **a** case is remarkable for **a** cooperation . . .			X
Coord.	I have also tried monthly data **and** the results are the same.			✓
	Rooms very clean **but** smelled very fresh.			X
EOS	the forehead is gathered in a frown **//** the mouth is slightly parted to reveal the teeth			✓
	the forehead is gathered in a frown the mouth **//** is slightly parted to reveal the teeth			X
	NLI			
Prep.	**With** a single jerk the man's head tore free.	→	The man's head tore free **from** a single jerk.	✓
	With a single jerk the man's head tore free.	→	The man's head tore free **without** a single jerk.	X
Negation	This is a common problem.	→	This is **not** an **uncommon** issue we are facing.	✓
	This is **not** a common problem.	→	This is **not** an **uncommon** issue we are facing.	X
Spatial	To reach . . . turn **left** up a small alleyway	→	do not turn **right** up the alleyway . . .	✓
	To reach . . . turn **left** up a small alleyway	→	Turn **right** up the alleyway . . .	X
Quant.	**all** taken up yeah	→	There are not still **some** left	✓
	all taken up yeah	→	There are still **some** left	X
Comp.	Today there are **more** than 300,000.	→	Today there are not **less** than 300,000.	✓
	Today there are **more** than 300,000.	→	Today there are **less** than 300,000.	X

Table 1: Examples of sentences and sentence pairs corresponding to each of our probing datasets. The highlighted words are those that are relevant to the phenomena targeted by each set.

holding the model architecture constant.

- We study function words, which have been under-studied in previous works on representation learning, but are critical to language understanding.
- We release nine new datasets (anonymized link), quality-controlled by both linguists and non-linguist annotators, to facilitate ongoing work and follow-up analysis.

2 Function Word Probing Tasks

2.1 Approach

We introduce nine new probing tasks aimed at evaluating models' understanding of function words. We focus on function words because although they are key building blocks of compositional meaning and are highly frequent, they have received relatively little attention in the probing literature and in the literature on distributional semantics. Each task targets the understanding of a specific type of function word; examples from each task are given in Table 1. Our expectation is that different pretraining objectives (see Section 3.2) will yield sentence representations which measurably differ in their performance on these probing tasks.

We use two different formats for our probing tasks: acceptability judgment and natural language inference (NLI). The former uses a binary classification approach (acceptable/unacceptable) for probing a single sentence vector, in line with works such as Conneau et al. (2018) and Adi et al. (2017). The latter uses an entailment-based approach similar to White et al. (2017) and Poliak et al. (2018), which is a ternary classification task (*entailment*, *contradiction*, *neutral*) over sentence pairs. The format is selected based on the suitability to the particular function word type in question.

To generate our probing datasets, we make structural modifications to sentences drawn from existing corpora, targeting a particular type of function word. We heuristically apply modifications which we believe are likely to produce a specific label, and then recruit human annotators in order to produce the final labels used in our evaluations. The result is a publicly available suite of nine task datasets (four acceptability tasks and five NLI tasks) consisting of 3,710 annotated examples. Appendix E lists the sizes of each dataset.

2.2 Acceptability Judgment-Based Tasks

We cast acceptability as a binary classification task following the format of such judgments commonly used in linguistics, in a similar manner to Warstadt et al. (2018). All tasks follow a common protocol of first identifying sentences that contain the construction that we are interested in, and then mutating half of the identified sentences to generate infelicitous versions of the original sentences. Unless stated otherwise, the original sentences are drawn from the test set of the Billion Word Benchmark (BWB, Chelba et al., 2013).

Wh-Words Understanding *wh*-words (i.e. *who, what, where, when, why, how*) depends on understanding the context and correctly identifying the antecedent, which may not be overtly present in the sentence; for instance, recognizing the infelicity of *I talked about who I live* requires knowing that the (unstated) antecedent must be a place and not a person. Our dataset consists of sentences that contain one of the six *wh*-words listed above. Half of these sentences are mutated versions of the original which are generated by replacing the original *wh*-word with a different *wh*-word randomly selected from the remaining five options.

Definite-Indefinite Articles The definiteness task probes the understanding of definiteness that arises by the use of the definite article (*the*) versus indefinite articles (*a* and *an*). We find sentences containing multiple occurrences of *the* or multiple occurrences of *a*, and, for half of them, swap all such occurrences (i.e. replacing *the* with *a*[1] or vice-versa). This gives us four types of sentences: unchanged sentences with multiple definite articles, unchanged sentences with multiple indefinite articles, sentences with all definite articles replaced by the indefinite article, and sentences with all indefinite articles replaced by the definite article. Our intent is that the former two types will be judged felicitous while the latter two will be infelicitous despite the fact that the sentence would be syntactically well-formed. We only focus on the cases with multiple occurrences of the same article, because replacing a single article most of the time did not significantly affect the acceptability (although it often did affect the actual meaning).

Coordinating Conjunctions Correct understanding of coordinating conjunctions (*and, but, or*) requires contextual comprehension of the two conjoined linguistic units, since different coordinating conjunctions express different logical relations, meaning their use is often restricted by the meanings of the conjoined items. We take sentences that contain coordinating conjunctions, and replace half of them with a version that contains a different conjunction. For example, the sentence *Room's very clean* ***but*** *smelled very fresh* is infelicitous despite being syntactically well-formed; *but* is unnatural here because the conjoined clauses do not form a clear contrast. Judging this sentence to be infelicitous requires a proper understanding of the ideas expressed in the clauses and how they relate to each other.

End-of-Sentence The end-of-sentence (EOS) task tests a model's ability to identify semantically coherent chunks (i.e., sentences) in running text. In written text this is often indicated by punctuation marks such as periods, but humans are able to easily identify sentences even without overt markers. Thus, we take pairs of sentences from the same paragraph of the WikiText-103 (Merity et al., 2017) test set and remove all punctuation marks and capitalization, and concatenate each sentence pair to create a line of running text.[2] Half of the dataset consists of a pair of valid sentences, and the other half consists of a pair of potentially invalid sentences generated from an incorrect segmentation of the running text, where the incorrect segmentation index is obtained by sampling from a Gaussian distribution centered around the correct index ($\sigma = 2$) and rounding to the nearest integer.

2.3 NLI-Based Tasks

Our NLI-based probing tasks ask whether the choice of function word affects the inferences licensed by a sentence. These tasks consist of a pair of sentences—a premise p and a hypothesis h—and ask whether or not p entails h. We exploit the label changes induced by a targeted mutation of the sentence pairs taken from the Multi-genre Natural Language Inference dataset (MNLI, Williams et al., 2018). The rationale is that, if a change to a single function word in the premise changes the entailment label, that function word must play a significant role in the semantics of the sentence.

[1] When we replace *the* with *a*, we choose *a/an* as necessary based on the word it precedes.

[2] We use WikiText instead of BWB because adjacent sentences in BWB are not logically contiguous and therefore may not be from the same discourse context.

Prepositions We manually curate a list of prepositions (see Appendix F) that are likely to be swapped with each other without affecting the grammaticality of the sentence. We generate mutated NLI pairs by finding occurrences of the prepositions in our list and randomly replacing them with other prepositions in the list. Our list consists of a set of locatives and several other manually-selected prepositions that are not strictly locatives but are likely to be substitutable (*about*, *for*, *to*, *with*, *without*).

Comparatives Comparatives assert relations among their arguments. For instance, a sentence that states *A is more than B* and another that states *B is more than A* lead to different inferences. We select a list of common comparatives (e.g., *more/less, bigger/smaller*) and select pairs from MNLI that contain a comparative phrase in both the premise and the hypothesis. We apply several mutations to the sentences, including negating the premise and/or hypothesis, and swapping comparatives (e.g. replacing *bigger* with *smaller*).[3]

Quantification The quantification task tests the understanding of natural language expressions of quantities, including common quantifiers (*all, some*), number words (*two, twenty*), and proportion (*half, one-third, quarter*). We select NLI pairs that contain at least one quantifier in both the premise and the hypothesis, and apply mutations of negating sentences and/or replacing quantifiers with syntactically appropriate substitutes.

Spatial Expressions The spatial expressions task probes the understanding of words that denote spatial relations between entities. Changing the spatial configuration often leads to different inferences; for instance, *A is to the left of B* implies that *B is to the right of A*, but not that *A is to the right of B*. We select a set of words that describe spatial configurations which are not necessarily prepositions (e.g., *left, right, close, far*). Again, we find MNLI pairs containing these words and negate/substitute to generate mutated pairs.

Negation This task probes whether models are able to understand negations, in particular explicit negation using the word *not*, lexical negation using antonyms, and the interaction between them. We first identify premise-hypothesis pairs from the MNLI dataset that contain antonym pairs (e.g., *dirty* appears in p and *clean* in h) and generate all possible patterns of negation with the two mutation strategies: swapping antonyms and adding explicit negation. That is, we use each of lexical negation, explicit negation, and their combination to mutate the premise and/or the hypothesis.

[3]We note that Dasgupta et al. (2018) also focus on comparatives, but they exclusively look at artificial sentences containing *more/less*.

2.4 Annotation

We recruit human annotators on Amazon Mechanical Turk to produce the final labels for the heuristically-generated datasets just described. We collect three labels per sentence (or per pair of sentences for EOS and NLI probing sets). We use the majority label in our final dataset, and discard examples on which there is no majority consensus. For more details about our annotation protocol, including compensation, refer to Appendix E.

Acceptability Tasks Human annotators are presented with a single (mutated or unmutated) sentence and are given the options {*natural*, *unnatural*, *neither*}. We discard sentences in which the majority label does not agree with our expected label. That is, we only include mutated sentences with a majority label of *unnatural* and unmutated sentences with a majority label of *natural*. We collect around 500 annotated examples with balanced label ratio for each probing set. We release our sentences in small batches until we have approximately 250 *unnatural* examples per task. To create the final dataset, we pool all answers from all batches and take a subset of the *natural* sentences so that the label ratio is balanced, prioritizing examples with perfect inter-annotator agreement.

Natural Language Inference Tasks For the NLI tasks, we collect common sense entailment judgments from human annotators on a 5-point Likert scale on which 1 denotes 'definitely contradiction' and 5 denotes 'definitely entailment', following Zhang et al. (2017). This finer-grained scale is intended to avoid confounds arising from borderline cases. Except for the use of scaled judgments, our instructions follow the MNLI guidelines. Specifically, our instructions said to assume that the sentences co-refer and that the first sentence (p) states a true fact, describes a scenario, or expresses an opinion, and to then indicate how likely it is that the second sentence (h) is also true, describes the same scenario, or expresses the same opinion. Annotators could also select an option indicating that one or both of the sentences did not

make sense; we discarded (p, h) pairs for which at least one annotator chose this option. We map judgments of 5 and 4 to *entailment*, 3 to *neutral*, and 2 and 1 to *contradiction*, and treat the majority label as the correct label after this mapping.

Agreement and Quality Control In constructing our final evaluation sets, we removed examples on which there was no majority consensus. For the binary acceptability tasks, we manually prefiltered sentences that were felicitous even after the heuristic modification. For the NLI tasks, we removed pairs that contained ungrammatical sentences that were not flagged by annotators via manual postfiltering. See Appendix E for more details.

3 Experimental Design

3.1 Pretraining Architecture

Since our focus is on comparing differences in pretraining objectives, we fix the architecture for all sentence encoders. We use the pretrained character-level convolutional neural network (CNN) from ELMo (Peters et al., 2018) that replaces word embeddings (see Bowman et al. (2018) or Tenney et al. (2019) for similar usages of the CNN layer). This acts as a base input layer that uses no information beyond the word, and allows us to avoid potentially difficult issues surrounding unknown word handling in transfer learning.

We feed the word representations to a 2-layer 1024d bidirectional LSTM (Hochreiter and Schmidhuber, 1997). A downstream task-specific model sees both the top-layer hidden states of this model and, through a skip connection, the original representation of each word. We train a version of this model on each task in Section 3.2. Additional model details are in Appendix A. Our codebase is open-source[4] and built using AllenNLP (Gardner et al., 2017) and PyTorch (Paszke et al., 2017).

3.2 Pretraining Tasks

Our main experiments compare seven pretraining tasks which we believe capture different aspects of linguistic meaning and which yield reasonable performance when used on a benchmark task such as MNLI.[5] For our purposes, a *task* is a dataset-training objective pair. We attempt to select a set of tasks diverse enough to highlight performance differences due to pretraining objectives. We additionally report results using BERT (Devlin et al., 2019) (base, uncased) to demonstrate that our probing sets prove challenging even for state-of-the-art models.

Language Modeling We train a left-to-right word-level language model on BWB, which was successfully used by Peters et al. (2018) for pretraining sentence encoders.

Skip-Thought Drawing from Kiros et al. (2015) and Tang et al. (2017), we train a sequence-to-sequence model on skip-thought, which is a task of generating the next sentence in the discourse given the previous sentence. We use the learned encoder as our sentence encoder. Since this objective requires running text, we use sentences from WikiText-103 as training data.

CCG Supertagging We train a model to predict the Combinatory Categorial Grammar (CCG) supertag for each word, with sentences from CCGBank (Hockenmaier and Steedman, 2007). Supertags are similar to part-of-speech tags but capture more syntactic context ("almost-parsing"; Bangalore and Joshi (1999)).

Discourse (DisSent) We train a model on DisSent (Jernite et al., 2017; Nie et al., 2017), which is an unsupervised task of predicting the discourse marker (e.g., *and*, *because*, or *so*) that connects two clauses. We train our model on a dataset created from WikiText-103 following Nie et al. (2017)'s protocol, which involves extracting pairs of clauses with a specific dependency relation.

Natural Language Inference Inspired by Conneau et al. (2017), we use the MNLI dataset for NLI pretraining. The task is to predict the entailment label for premise-hypothesis pairs; the possible labels are *entailment, contradiction, neutral*.

Machine Translation We train a sequence-to-sequence machine translation model with attention on WMT14 English-German (Bojar et al., 2014) and take the encoder as our sentence encoder. McCann et al. (2017) previously showed that pretraining an encoder on translation led to good performance on downstream NLP tasks.

Image-Caption Matching We train a model on the task of grounding sentences to the images they describe. We use image-caption pairs from the MSCOCO dataset (Lin et al., 2014) with an objective that minimizes the cosine distance between

[4] Link anonymized for submission.

[5] Around 70% accuracy; see Appendix D for details.

sentence representations and corresponding image features, as described in Kiela et al. (2018).

3.3 Task-Specific Classifiers

To probe the sentence encoders pretrained on the different objectives, we freeze the weights of the encoder after pretraining and train an additional model using the outputs of the fixed encoder as inputs. We describe the implementation details for the NLI and acceptability probing sets below.

NLI Tasks For the NLI-type probing sets, we train an NLI model on top of the representations produced by the pretrained sentence encoder that uses an attention mechanism inspired by Seo et al. (2017) that computes attention between all pairs of words in the two sentences (see Appendix A for more implementational details). We train this component on MNLI and evaluate directly on our NLI probing sets with no further training.

Acceptability Classification Tasks For all acceptability tasks except the EOS task, we take the sequence of hidden state outputs from the pretrained encoder as the sentence representation. We aggregate this sequence into a single vector via max-pooling and train a 512d MLP on top of the resulting vector. For the EOS task, we also use max-pooling on each sentence in the pair. We then concatenate the resulting vectors and train an MLP on top of the joint representation.[6] Each task has around 400 training examples (see Appendix E). Due to their small size, we use 10-fold cross validation where each fold is used as the test set exactly once, and report the average test set accuracy.

3.4 Variation from Random Restarts

In order to calibrate the degree of variation that can be expected due to random noise, we run each of our probing tasks on five different random initializations of the sentence encoder weights. These sentence encoders were not pretrained, and we trained MLPs for each probing task on top of the randomly initialized sentence encoders. Across five random restarts, the average standard deviation across our probing set was around 1 percentage point. The mean and 95% confidence interval for each probing task are reported in Appendix G.

4 Results

4.1 Overall Performance

Figure 1 shows the performances of models trained on each pretraining task on our probing datasets. We observe that different pretraining tasks have different strengths and weaknesses. There is no single pretraining task that achieves the best (or worst) performance across the board. This implies that even the best encoders, such as BERT, are unable to capture function word semantics fully, and suggests further research into combining advantages of different tasks. Furthermore, most models are far from human performance, with only a few exceptions (e.g., BERT on conjunctions). This demonstrates that our probing datasets serve as useful challenge sets, in addition to permitting fine-grained probing analysis.

Looking into each probing set in more detail, we see several intuitive patterns on how pretraining might affect performance on the probing sets. Among the pretrained models (not including BERT), the NLI model did best on the negation[7] and conjunction tasks, both of which involve words that play central roles in inferential reasoning. The CCG model yields the best result for EOS, which could be attributed to the task's emphasis on structure; it is the only task that imposes a restriction on structural composition.

Surprisingly, we find that pretraining can sometimes hurt performance. For instance, pretraining uniformly hurts performance on comparatives with the exception of skip-thought, which is still within random variation range. In fact, for many probing sets, the choice of pretraining task affects whether it helps or hurts performance; for instance, pretraining on NLI helps with negation, whereas pretraining on image-caption matching and CCG lowers performance. This suggests that pretraining can be helpful, but *only* helpful if we pretrain on a task that provides useful information in solving the probing set. For instance, in Section 4.3 we discuss how the image-caption matching objective may bias models to discard information about certain preposition senses. Overall, we

[6] We tried training a general acceptability model using CoLA and evaluating directly on our acceptability tasks, as an analogous evaluation setup to the NLI tasks, but all models performed around chance under this setup. This is likely due to the intrinsic difficulty of CoLA for our base model, as suggested by low performance from similar models ("GLUE Baselines") on gluebenchmark.com.

[7] We additionally find that this improvement is specifically due to the NLI model's capacity to understand explicit negation using *not*, rather than lexical negation with antonymy. See Appendix H for differences between negation subtypes.

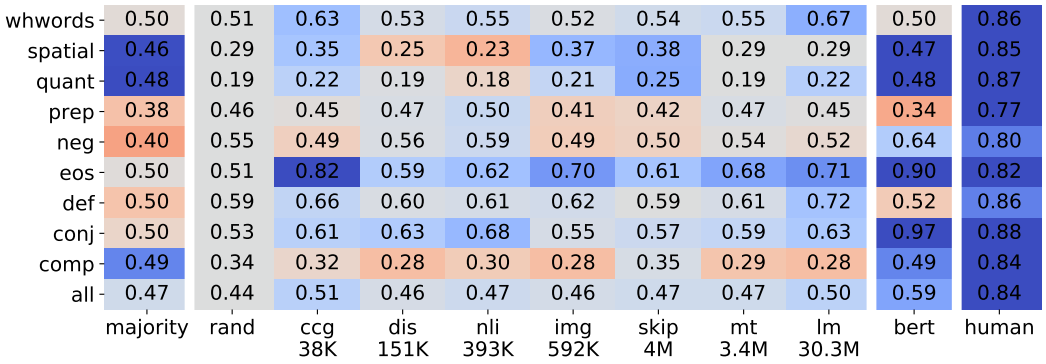

Figure 1: Accuracy for each pretraining task on each probing set. The leftmost column shows the majority-class baseline, and the rightmost column shows individual annotator accuracy on the final probing set. Blue denotes performance improvement over randomly initialized encoder baseline and orange denotes performance decrease.

observe that language modeling is a useful pretraining task, which aligns with its effectiveness for pretraining models that achieve state-of-the-art NLP results. However, the most beneficial task on average (in terms of both raw accuracy and gains over random baseline) is CCG, our most syntactic task, which suggests that syntactic knowledge helps with function word comprehension.

We also see that our probing sets are challenging even for BERT—although BERT substantially improves performance on many probing sets, and obtains superhuman performance on conjunctions and EOS,[8] it also shows clear weaknesses in several probing sets (e.g. *wh*-words and prepositions) where it is outperformed even by a randomly initialized baseline with no pretraining.

4.2 Correlations between Pretaining Tasks

To further investigate whether our probing sets differentiate between pretraining objectives, we look into correlations between the model predictions; given two pretraining tasks i and j, how often does a model trained on i make the exact same prediction as a model trained on j? Figure 2 shows the correlations across all probing sets in aggregate, and for the *wh*-words and prepositions sets specifically (see Appendix I for all sets).

We observe that models pretrained on different tasks do make different predictions overall, with image-caption matching and skip-thought being the tasks that make predictions that deviate the most from others (left). NLI and image-caption matching are the least correlated pair of tasks among all. The difference between image-captioning and other tasks is the most prominent in the preposition probing set; it makes predictions that are only weakly correlated with others (middle). We hypothesize that this is due to the duality of preposition semantics; most prepositions have both concrete and abstract senses, and the image model is biased to focus on the former.

To illustrate, consider the preposition *below*, which can denote a spatial configuration (e.g. *the boots end* ***below*** *the knee*) or an abstract relation (numeric or qualitative comparison; e.g. *her height is* ***below*** *six feet*). In the preposition dataset, *below* occurs 17 times, 11 of which are spatial and 6 abstract. For the spatial usage, both MNLI and image-caption models have 64% accuracy (7/11). The NLI model shows 50% accuracy for pairs containing abstract uses (3/6), but the image-captioning model answers none of them correctly (0/6). Here is an example of a numeric usage of *below* that the NLI model answered correctly but the image model answered incorrectly:

P: *Only those whose incomes do not exceed 125 percent of the federal poverty level qualify . . .*

H: *Those whose incomes are* ***below*** *125 percent qualify . . .* (P→H)

The image model's bias towards the spatial usage is intuitive, since the numeric usage of *below* (i.e. as a counterpart to *exceed*) is difficult to learn from visual clues only. This concrete-abstract duality, which is not specific to *below* but common to most other prepositions (Schneider et al., 2018), may partially explain why the image-caption model behaves so differently from all other models, which are not trained on a multimodal objective.

[8] We speculate that this might be an effect of the next-sentence classification task that BERT is pretrained on.

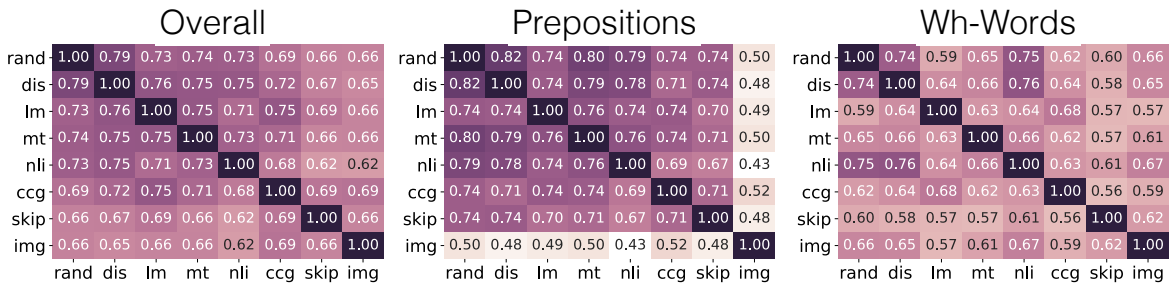

Figure 2: Prediction overlap on the probing tasks for models trained on different pretraining tasks (i.e. how often models make identical predictions on a particular probing set).

4.3 Data Size and Genre Effects

As can be seen from the varying sizes of the pretraining dataset reported in Table 1, seeing more data at pretrain time does not imply better performance on probing tasks. Also, as noted before, the fact that pretraining can hurt probing performance suggests that if the task is not the "right" task, adding more datapoints at pretrain time is not necessarily beneficial for probing performance.

Another potential confound is vocabulary overlap between pretraining and probing task datasets. Since all pretraining task datasets have different sets of vocabulary, the variance in the results could be attributed to the amount of words in the probing set already seen at pretrain time. To investigate this possibility, we compute the ratio of overlapping words between the pretraining and probing datasets. A regression analysis shows that vocabulary overlap overall does not predict better performance on the probing set ($p > .05$). No single probing set performance was significantly affected by vocabulary overlap either (all $p > .05$ after Bonferroni correction for multiple comparisons).

5 Related Work

An active line of work focuses on "probing" neural representations of language. Ettinger et al. (2016); Ribeiro et al. (2018); Zhu et al. (2018); Naik et al. (2018), *i.a.,* use a task-based approach to probing similar to ours, where tasks that require a specific subset of linguistic knowledge are used to perform qualitative evaluation. Gulordava et al. (2018), Giulianelli et al. (2018), Rønning et al. (2018), and Jumelet and Hupkes (2018) make a focused contribution for a particular linguistic phenomenon (e.g., agreement, ellipsis). Staliūnaite and Bonfil (2017) and Mahler et al. (2017) use a similar strategy to our structural mutation protocol, although their focus was on breaking existing systems rather than comparing different models.

The design of our NLI-style probing tasks follows the recent line of work which advocates for NLI as a general-purpose format for diagnostic tasks (White et al., 2017; Poliak et al., 2018). This idea is similar in spirit to McCann et al. (2018), which advocates for question answering as a general-purpose format, to edge probing (Tenney et al., 2019) which probes for syntactic and semantic structures via a common labeling format, and to GLUE (Wang et al., 2018) which aggregates a variety of tasks that share a common sentence-classification format. The primary difference in our work is that we focus specifically on the understanding of function words in context. We also present a suite of several tasks, but each one focuses on a particular structure, whereas works above generally aggregate multiple phenomena.

6 Conclusion

We propose a new challenge set of nine tasks that focus on probing function word comprehension. Although we use our challenge set to compare the effects of pretraining, the probing sets themselves are architecture- and evaluation setup-agnostic. The results show that models pretrained with different objectives do generate different predictions (e.g., image models have a bias towards concrete preposition senses), and that no single objective leads to models that perform best or worst across all probing tasks. This suggests that there are 'gaps' in the linguistic knowledge learned from a single pretraining objective that could be complemented by other objectives, and this calls for further research into how different pretraining objectives could be productively combined. Additionally, we hope that our exploratory study initiates further discussions about modeling function words and their contribution to compositional meaning.

Acknowledgments

The work reported here was conducted at the 2018 Frederick Jelinek Memorial Summer Workshop on Speech and Language Technologies, and supported by Johns Hopkins University with unrestricted gifts from Amazon, Facebook, Google, Microsoft and Mitsubishi Electric Research Laboratories

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *International Conference on Learning Representations*.

Srinivas Bangalore and Aravind K Joshi. 1999. Supertagging: An approach to almost parsing. *Computational linguistics*, 25(2):237–265.

Raffaella Bernardi, Gemma Boleda, Raquel Fernandez, and Denis Paperno. 2015. Distributional semantics in use. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 95–101, Lisbon, Portugal. Association for Computational Linguistics.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Samuel Bowman, Yoav Goldberg, Felix Hill, Angeliki Lazaridou, Omer Levy, Roi Reichart, and Anders Sgaard, editors. 2017. *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*. Association for Computational Linguistics, Copenhagen, Denmark.

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.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv:1312.3005*.

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

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136. Association for Computational Linguistics.

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

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 Papers)*.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. AllenNLP: A deep semantic natural language processing platform.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. 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 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Julia Hockenmaier and Mark Steedman. 2007. Ccgbank: A corpus of ccg derivations and dependency structures extracted from the penn treebank. *Computational Linguistics*, 33(3).

Yacine Jernite, Samuel R Bowman, and David Sontag. 2017. Discourse-based objectives for fast unsupervised sentence representation learning. *arXiv:1705.00557*.

Jaap Jumelet and Dieuwke Hupkes. 2018. Do language models understand anything? O n the ability of lstms to understand negative polarity items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 222–231, Brussels, Belgium. Association for Computational Linguistics.

Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. 2018. Learning visually grounded sentence representations. 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 408–418, New Orleans, Louisiana. Association for Computational Linguistics.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, pages 3294–3302.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer.

Tal Linzen, Grzegorz Chrupala, and Afra Alishahi, editors. 2018. *Proceedings of the First Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP)*. Association for Computational Linguistics, Brussels, Belgium.

Taylor Mahler, Willy Cheung, Micha Elsner, David King, Marie-Catherine de Marneffe, Cory Shain, Symon Stevens-Guille, and Michael White. 2017. Breaking nlp: Using morphosyntax, semantics, pragmatics and world knowledge to fool sentiment analysis systems. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 33–39. Association for Computational Linguistics.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6297–6308.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv:1806.08730*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353.

Allen Nie, Erin D. Bennett, and Noah D. Goodman. 2017. DisSent: Sentence representation learning from explicit discourse relations. *arXiv:1710.04334*.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. *NIPS 2017 Autodiff Workshop*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237. Association for Computational Linguistics.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865. Association for Computational Linguistics.

Ola Rønning, Daniel Hardt, and Anders Søgaard. 2018. Linguistic representations in multi-task neural networks for ellipsis resolution. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 66–73. Association for Computational Linguistics.

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Jakob Prange, Austin Blodgett, Sarah R. Moeller, Aviram Stern, Adi Bitan, and Omri Abend. 2018. Comprehensive supersense disambiguation of english prepositions and possessives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 185–196. Association for Computational Linguistics.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. *International Conference on Learning Representations*.

Ieva Staliūnaite and Ben Bonfil. 2017. Breaking sentiment analysis of movie reviews. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 61–64. Association for Computational Linguistics.

Shuai Tang, Hailin Jin, Chen Fang, Zhaowen Wang, and Virginia de Sa. 2017. Rethinking skip-thought: A neighborhood based approach. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 211–218. Association for Computational Linguistics.

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, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv:1804.07461*.

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

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005. Asian Federation of Natural Language Processing.

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. Association for Computational Linguistics.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association of Computational Linguistics*, 5(1):379–395.

Xunjie Zhu, Tingfeng Li, and Gerard de Melo. 2018. Exploring semantic properties of sentence embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 632–637, Melbourne, Australia. Association for Computational Linguistics.