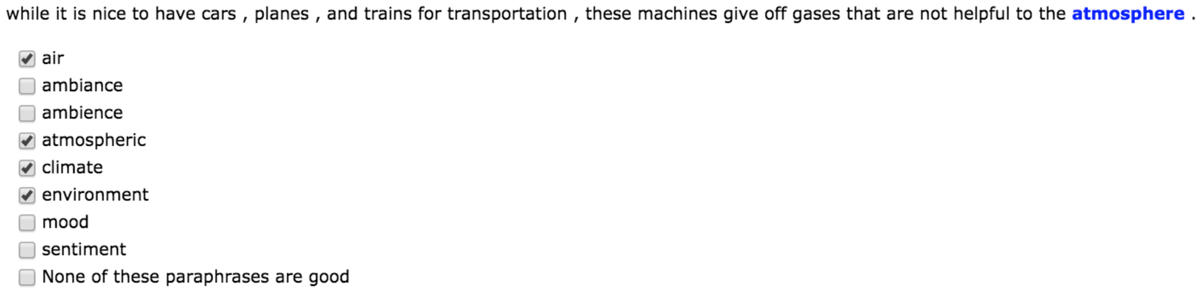
Domain-Specific Paraphrase Extraction: Supplementary Material

1 Annotation methodology for evaluation data

We want to collect labeled data to allow us to evaluate each of our paraphrase models in terms of the precision and recall of $\langle e_1, e_2 \rangle$ pairs. To do this, we collect human judgements of paraphrase appropriateness in the target domain. We take a sample of 15K sentences from our biology data. We select a phrase (e_1) from each sentence and gather judgements for all of its candidate paraphrases (e_2s). These candidates are taken from the general (non-domain-specific) paraphrase model, and constitute the full set of paraphrases that can be extracted from our training corpus. We show the sentence, the original phrase, and the full list of candidate paraphrases to 5 workers on Amazon Mechanical Turk. The workers make a binary judgement for whether each paraphrase is appropriate given the sentence context (see Figure 1).

We consider e_2 to be a good paraphrase for e_1 in the domain if it was judged to be good in least one sentence by the majority of workers. We collect analogous judgements for 10K sentences from the general domain, chosen randomly from Wikipedia, which we use as a control. Table 1 shows some examples of paraphrases (and the associated sentences) that were judged to be good and bad in the biology and general domain.

To ensure good quality, we embedded quality control questions using WordNet synonyms and synset-specific example sentences. Each quality control question consisted of at least three good paraphrases (which we expected workers to select) and three randomly chosen words (which we expected workers not to select). We rejected workers who fell below 50% accuracy after completing 10 or more HITs. Overall, worker accuracy was 82% and inner-annotator agreement was κ=0.65.



while it is nice to have cars , planes , and trains for transportation , these machines give off gases that are not helpful to the **atmosphere** .

☑ air
□ ambiance
□ ambience
☑ atmospheric
☑ climate
☑ environment
□ mood
□ sentiment
□ None of these paraphrases are good

Figure 1: HIT interface shown to workers on MTurk

Paraphrase pair	Label	Domain	Context
plant/factory	Good	General	games were played at traktor stadium , which is located near the **plant** .
plant/factory	Bad	Biology	the functions of a plant 's roots are to support the **plant** and make food .
ground/earth	Bad	General	that show never got off the **ground** .
ground/earth	Good	Biology	a plant that lives in the desert might have a large root system to find water deep within the **ground** .

Table 1: Majority vote from workers for whether a paraphrase is appropriate given a general domain context and a domain-specific context. The final binary labels for a given paraphrase pair were aggregated across all contexts. I.e. a pair is considered "good" for a domain if it was "good" in at least one context according to the majority of workers.