A large annotated corpus for learning natural language inference

The paper "A large annotated corpus for learning natural language inference" by Samuel R. Bowman, Gabor Angeli, Christopher Potts and Christopher D. Manning, published in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* in 2015, introduced Stanford Natural Language Inference (SNLI) corpus, a collection of 570K labeled sentence pairs for understanding entailment and contradiction, to provide important resource for machine learning in natural language inference.

Before SNLI, the natural language inference corpora are generally too small or training modern data-intensive, wide-coverage models, many contain sentences that were algorithmically generated, and they are often beset with indeterminacies of event and entity coreference that significantly impact annotation quality. To address the issues of size, quality, and indeterminacy, the authors employed a crowdsourcing framework with the following crucial innovations. First, the examples were grounded in specific scenarios, and the premise and hypothesis sentences in each example were constrained to describe that scenario from the same perspective, which helps greatly in controlling event and entity coreference. Second, the prompt gave participants the freedom to produce entirely novel sentences within the task setting, which led to richer examples than we see with the more proscribed string-editing techniques of earlier approaches, without sacrificing consistency. Third, a subset of the resulting sentences were sent to a validation task aimed at providing a highly reliable set of annotations over the same data, and at identifying areas of inferential uncertainty. The corpus was distributed with a pre-specified train/test/development split. The test and development sets contain 10k examples each. The distributed corpus includes parses produced by the Stanford PCFG Parser 3.5.2, trained on the standard training set as well as on the Brown Corpus.

Testing the corpus on models, the author found that the large size of the corpus is crucial to both the LSTM and the lexicalized model, and suggests that additional data would yield still better performance for both. In addition, though the LSTM

and the lexicalized model show similar performance when trained on the cur- rent full corpus, the somewhat steeper slope for the LSTM hints that its ability to learn arbitrarily structured representations of sentence meaning may give it an advantage over the more con- strained lexicalized model on still larger datasets. Also, they evaluate on the SICK entailment task using a simple transfer learning method and achieve competitive results. To perform transfer, they took the parameters of the LSTM RNN model trained on SNLI and use them to initialize a new model, which was trained from that point only on the training portion of SICK. The only newly initialized parameters were softmax layer parameters and the embeddings for words that appear in SICK, but not in SNLI. Transferring SNLI representations to SICK yields the best performance yet reported for an unaugmented neural network model, surpasses the available EOP models, and approaches both the overall state of the art at 84.6% and the 84% level of interannotator agreement, which likely represents an approximate performance ceiling. This suggests that the introduction of a large high-quality corpus makes it possible to train representation-learning models for sentence meaning that are competitive with the best hand-engineered models on inference tasks.