Literature Review: Natural Language Inference xCS224U - April 2020
Alan Dunetz

General Problem/Task Definition: what are these papers trying to solve and why?

Recognizing Textual Enhancement (RTE) is the task of determining whether for two sentence:
i. the first sentence (the premise) entails the second sentence;
ii. the two sentences are unrelated; or
iii. the two sentences are contradictory.

The RTE task is intended to show whether a computer model understands the semantics/ meaning of sentences. It is therefore a key challenge in the field of natural language understanding.

The first three papers describe key advances in models and datasets used to test RTE. The remaining two papers discuss issues and limitations with these models and suggest some possible approaches to address these limitations.

Concise Summaries

1. Dagan, et. al., The PASCAL Recognising Textual Entailment Challenge
The paper describes the Recognizing Textual Entailment (RTE) Challenge Benchmark. The textual entailment task is defined as a directional relationship between pairs of text fragments: "T", the entailing text; and "H" the hypothesis. The task is to evaluate whether the hypothesis can be inferred from the entailing text. The authors created a dataset of T-H pairs of text snippets. Human annotators manually labelled each pair for entailment and the dataset was divided into developmental and test datasets. Each system tested had to determine, for each pair, whether T entails H and the results were compared to the manual standard.

The dataset was biased towards non-trivial examples and the true and false label categories were approximately balanced. The T portion of each pair typically consisted of one sentence while H was a shorter sentence. The pairs were collected from external sources including available datasets and the web. Annotators decided which pairs to include. The authors note, that, in retrospect, the annotators selected more negative examples than positive ones in the case where there was a high degree of lexical overlap between T and H.

The dataset consisted of seven application sub settings including: Information Retrieval, (IR) Comparable Documents (CD), Reading Comprehension (RC), Question Answering (QA), Information Extraction (IE), Machine Translation (MT) and Paraphrase Acquisition (PP).

Each example was first judged by the annotator who created the example and then cross-evaluated by a second evaluator. The annotators agreed in about 80% of the cases. The remaining 20% were discarded. The authors eliminated an additional 13% of examples which they thought might be controversial. The final dataset included 567 examples in the development set and 800 in the test set.

Submitted systems tagged each T-H pair and optionally added a confidence score between 0 and 1. All systems were evaluated with respect to their accuracy compared to the manual labels (gold standard). In addition, systems were evaluated according to a second measure, the Confidence-Weighted Score (CWS or average precision) which rewards systems that assign higher confidence scores to correct responses.

Sixteen submitted systems were evaluated. The methods used by these systems included word overlap, statistical lexical relationship, WordNet, syntactic matching, world knowledge and logical inference. Some systems uses as many as four of these methods.

Overall system accuracies were between 50 and 60 percent and CWS scores were between 0.50 and 0.70. A system that simply predicted uniformly true or false would have achieved 50%.

The authors note that their method did not favor specific recognition of positive entailment and that standard measures including precision, recall and f in terms of the positive examples would have been useful as additional measures. None of the systems performed significantly better than the f=0.67 baseline of a system which predicts true in every case.

2. Bowman et. al., A large annotated corpus for learning natural language inference

The paper introduces a new dataset for natural language inference, the Stanford Natural Language Inference (SNLI) corpus. The corpus includes 570K labeled sentence pairs, over 400 times the size of the Dagan et. al. dataset. The larger dataset provides sufficient data to train more complex models, including neural networks on the RTE task.

The premises for the dataset were captions from the FLickr30k corpus, a dataset of approximately 160k crowd-sourced captions corresponding to about 30k photos. For each caption they were shown, workers were instructed to create a hypothesis for each of the three labels: entailment, neutral and contradiction. Consequently, the dataset is balanced between the labels. An additional round of validation was performed on 10% of the data which showed a high overall level of agreement with the original labels (91%).

The authors evaluated a variety of NLI models on the corpus including rule-based systems, linguistic classifiers and neural networks and found the best results with two models. The first is a feature-rich linguistic classifier and the second is an LSTM model.

The linguistic classifier employed three lexicalized and three non-lexicalized feature types. The inclusion of lexicalized feature types resulted in a significant increase in accuracy.

The authors compared the performance of three neural network models. Each model produced a vector representation of each of the two sentences separately. The two resulting vectors were concatenated and used as input to a multi-layer perceptron (MLP) . Their MLP is simply a stack of three tanh layers (200d) with the top layer feeding a softmax classifier.

The authors used 300d GloVe vectors to create word embeddings for each word. Their baseline approach simply summed the vectors of each word in a sentence to create the overall vector representation for that sentence. In addition, they looked at two RNN models: a simple RNN model and an RNN LSTM model. In both cases, the premise and hypothesis sentences were encoded separately and then concatenated. The LSTM model achieved the best result of these with an accuracy of 84.8% on the training set and 77.6% on the test set.

The learning curves for the RNN LSTM and the lexicalized classifier showed that the size of the dataset was critical for both models, and that both models would benefit from an even larger dataset.

In addition, the authors explored whether the performance of the LSTM model on the NLI task would benefit from transfer learning. For this purpose, they took the model parameters from the LSTM model trained on the SNLI task, and used these parameters to initialize a new LSTM model. This second model was then trained on separate NLI task, the SICK dataset. They

found that the second model performed better on the SICK dataset than either: a) a model trained exclusively on the SICK dataset or b) the model trained solely on the SNLI dataset. Their results approached the best state of the art results for the dataset and suggest that the availability of a high quality corpus can make representation learning models, such as LSTMs, more competitive across a wide range of NLI tasks.

Rocktanshel et.al., Reasoning about Entailment with Neural Attention
In Bowman et.al, (above) an LSTM model was trained separately on the premise and the hypothesis and then the two vectors were concatenated and passed as input to a neural network classifier. Rocktanshel et. al. extends this analysis on the SNLI dataset in a number of ways. First, their LSTM model trains the premise and hypothesis together to determine entailment. The authors believe this allows information to flow from the premise to the hypothesis. Second, they extend their model to use an attention mechanism. Finally, they provide qualitative analysis which show which output representation their model attends to when it determines the class of a representation.

LSTMs contain memory cells that can store information long term. In addition, they contain three types of gates, input gates, forget gates and output gates which control the flow of information from one time step to the next. At each time step, the input to a memory cell is the current input vector x_t, the previous output vector h_{t-1} and the cell state vector c_{t-1}. The LSTM calculates the next output state vector, h_t, and the next cell state vector c_t

Their model uses word2vec vector representations instead of GloVe vectors. Out-of-vocabulary words are initialized randomly. A linear layer projects the word vectors to the dimensionality of the LSTMs.

In their model the premise is read by one LSTM model and the hypothesis, preceded by a delimiter, is read by a second LSTM model. But the second model is initialized with the final hidden state of the first model. The final hidden state of the second model is then input into a simple neural network consisting of a single nonlinear layer and a softmax layer. The target space of the softmax is the three class labels, entailment, neutral or contradiction. The model is trained using cross-entropy loss.

In addition, the authors extended the model to incorporate attention mechanisms. Under this approach, the output state (h_t) for each word in the premise is saved and used to reduce information bottlenecks when processing the hypothesis. The authors used two types of attentions mechanisms. In their basic approach only the final output LSTM output state vector attended to the earlier output state vectors from the premise. In their word-by-word approach, each output vector in the hypothesis attended the output states from the premise.

Finally, the authors also considered a two-way attention approach which allowed the individual words of the premise to attend over all of the words of the hypothesis by interchanging the two sentences. In this model two sentence pair representations were generated. These representations were concatenated and input into the neural classifier.

The authors found that processing the hypothesis conditioned on the premise gave an improvement on the test set of 3.3% in accuracy over processing the two sentences separately. Enabling their basic attention model resulted in a further improvement of 0.9%. Enabling word-by-word attention resulted in a further improvement of 1.2% versus the base attention model. Enabling two-way attention did not have a significant impact either for the base attention or the word-by-word attention models.

The authors also provided some visualizations of which outputs the model attended to when classifying an instance. They observed that, in the basic attention model, the model pays attention to words that are semantically coherent with the premise. The authors found that the word-by-word model could recognize cases where the hypothesis is simply a rewording of the premise and was able to resolve synonyms.

4. *Poliak et.al.,Hypothesis Only Baselines in Natural Language Inference*
Each instance in an NLI dataset typically consists of a premise and a hypothesis. Models are scored based on their ability to determine whether the premise entails the hypothesis. Poliak et. al., argues that any interesting NLI model must therefore depend on both the premise and the hypothesis. If a model which only has access to the hypotheses can perform well then it is only performing NLI to the extent that inference is based on prior background knowledge. Alternatively, such a model may be exploiting quirks or statistical irregularities in the dataset.

Their paper explores whether NLI datasets contain such statistical irregularities. They collect 10 datasets which they categorize into three groups. The Human Elicited category included datasets such as SNLI and MNLI where humans were given a context and generated a hypothesis. The Human Judged category includes datasets, such as SICK, where hypothesis and premises were automatically generated and humans supplied the labels. Finally, the Automatically Recast category includes datasets automatically generated for other NLP tasks and labeled using minimal human intervention.

Their neural model is a Bi-directional LSTM encoder which constructs sentence representations by max-pooling over its hidden states. The recast datasets were preprocessed using the NLTK tokenizer. All of the datasets were mapped to GloVe 300d vectors and input to the neural model. They optimized using SGD with at most 20 epochs of training.

The authors found that in six out of ten cases, their hypothesis-only model significantly outperformed the majority baseline. For example, on the SNLI dataset, their hypothesis-only model achieved an accuracy of 69% on the Test set compared to an accuracy of only 34% using the majority baseline.Their results suggest that there are exploitable statistical irregularities that can allow NLI models to perform well without actually making an inference from the premise.

The authors explored the nature of these statistical irregularities and identified a number of potential causes including specific words, grammar and lexical semantics.

They found specific words that were highly correlated with labels. For example, in the SNLI dataset, words implying contradiction or removal of agency (e.g., ‘sleeping’ and therefore unable to act) were highly correlated with contradiction.

For some other datasets, the authors found correlation between a sentence’s label and whether it was grammatically correct. In these cases, a hypothesis-only model might be able to make correct classifications based on syntax.

Finally, the authors found that in some cases the models could discriminate based on the lexical semantics of the sentence. For example, a model could learn that a correct use of ‘sentient’ is that ‘experts are sentient’ and an incorrect use is ‘The campaign is sentient’.

The authors suggest that researchers should develop new NLI datasets that exhibit less exploitable irregularities and that hypothesis-only baselines should replace majority baselines.

5. Gururangen et. al., Annotation Artifacts in Natural Language Inference Data
Similar to Poliak et. al., the authors note that hypothesis generated from a crowdsourcing process may contain artifacts that enable a model to classify a hypothesis without looking at the premise. Rather than performing inference in such cases, they argue that the model is simply exploiting statistical irregularities.

The authors evaluate the SNLI and MNLI datasets using a hypothesis-only model. They used an off-the-shelf text classifier that models text as a bag of words and bigrams. They trained only on the hypotheses in each dataset but still achieved results well in excess of the majority baseline. For example, similar to Poliak et. al., they find their hypothesis-only model has an accuracy of 67% on the SNLI dataset versus a majority class baseline of 34%.

The authors speculate that annotation artifacts in elicited datasets are common. For example workers may apply strategies to generate their hypothesis quickly. The authors attempt to identify some of these artifacts based on category label.

In the entailment category, the authors find more generic words such as 'animal' or 'instrument' that were probably selected to generalize over more specific premise words such as 'cat' or 'piano'. Also exact numbers were replaced with approximates, e.g., 'some', or 'at least'. In addition, explicit gender terms were replaced with gender-neutral terms, e.g., 'human' or 'person'.

For the neutral category, the authors found the addition of modifiers, such as 'tall' or 'sad' or were added to introduce information that is not obviously entailed by the premise. Superlatives were also added in the neutral category for this purpose.

For the contradictory category, the authors found that negative words were indicative of contradiction. Like Poliak, they note that 'sleeping' contradicts any activity.

Separately, the authors found that entailed sentences tended to be shorter than sentences belonging to the other two categories. The authors suggest that workers may create entailed sentences by simply removing words from the premise.

The authors then reevaluate several high performing NLI models on a subset of instances where their hypothesis-only classifier failed. The models they selected include the Decomposable Attention Model (DAM), the Enhanced Sequential Inference Model (ESIM) and the Densely Interactive Inference Model (DIIN). Each model was trained on the full SNLI or MNLI dataset and then tested on: 1) the full dataset; 2) an 'easy' subset (includes instances the hypothesis-only model classified correctly); and 3) a 'hard' subset' (includes instances the hypothesis-only model failed to classify correctly).

They found that each model's success is primarily due instances from the 'easy' subsets. For example, the DAM model had an accuracy of 84.7% on the full SNLI dataset, and an accuracy of 92.4% on the easy dataset, but an accuracy of only 69.4% on the 'hard' dataset. This suggests that the models are less successful at performing natural language inference than previously believed.

They note that the 'hard' datasets may not be artifact-free. In addition, they note that the presence of artifacts do not make an example incorrect. The problem is how these artifacts are distributed across the labels. Therefore, for new datasets, it may be a better strategy to try to distribute artifacts evenly across labels than to try to eliminate them.

Compare and Contrast

Dagan was an initial attempt to create a benchmark to capture major semantic inferences across applications. Bowman introduced the Stanford Natural Language Inference (SNLI) corpus, a much larger dataset which provided sufficient data to train neural networks. Bowman was able to achieve results on this dataset with an LSTM model comparable to the results achieved by feature based classifiers. Rocktaschel applied a more sophisticated LSTM model to the SNLI dataset resulting in a significant improvement in results over Bowman. These enhancements included: 1) initializing the LSTM for the hypothesis with the final output state of the premise; and 2) adding attention mechanisms.

The remaining two papers, analyze limitations with NLI models. Poliak found that a hypothesis-only model could achieve results substantially better than the majority class baseline. Poliak attributed this performance to statistical irregularities in the datasets examined and identifies a number of sources of such irregularities including word-label correlation, grammaticalness, and lexical semantics. Gurangan also reports strong results for hypothesis-only models. Gurangan then reevaluates NLI models on subsets of the SNLI and MNLI datasets and finds that much of the success of these models is attributable to instances that can be correctly classified by the hypothesis-only model.

Future Work: Make several suggestions for how the work can be extended

1. Use additional metrics in place of accuracy including, for example, macro-F.
2. Use of contextual embedding models, such as BERT, on the NLI inference task.
3. Use of contextual embedding models on the hypothesis-only task.
4. Hypothesis-only analysis of the single word entailment task from HW3.

References

1. Ido Dagan, Oren Glickman, Bernando Magnini, The PASCAL Recognising Textual Entailment Challenge. In Machine Learning Challenges pages 177-190.
https://u.cs.biu.ac.il/~nlp/RTE1/Proceedings/dagan_et_al.pdf

2. Samuel R. Bowman, Christopher Potts, Gabor Angeli and Christopher Manning, A large annotated corpus for learning natural language inference, 2015. In Proc of EMNLP
https://arxiv.org/abs/1508.05326

3. Tim Rocktaschel, Edward Grefenstette, Karl Moritz Herman, Tomas Kocisky and Phil Blumson, Reasoning about Entailment with Neural Attention, 2015. In ICLR 2016
https://arxiv.org/pdf/1509.06664.pdf

4. Adam Poliak, Jason Nardowsky, Aparajita Haldar, Rachel Rudinger and Benjamin Van Durme, Hypothesis Only Baselines in Natural Language Inference, 2018.In Proc of *SEM
https://www.aclweb.org/anthology/S18-2023.pdf

5. Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman and Noah A. Smith, Annotation Artifacts in Natural Language Inference Data, 2018 In Proc of NAACL.
https://arxiv.org/abs/1803.02324