# An Automated and Exhaustive Natural Language Inference Corpus

Atticus Geiger and Michelle Tran
atticusg@stanford.edu, mchellet@stanford.edu

## Natural Language Inference

Natural language inference (NLI), also known as recognizing textual entailment (RTE), is the problem of determining whether a natural language hypothesis follows from or is contradicted by a natural language premise. A model that performs natural language inference takes in a premise and hypothesis sentence and then classifies the pair as either an entailment, contradiction, or permission.

A man happily did not kick the ball.
entails
The ball was not kicked by the man.

## Existing NLI Corpora

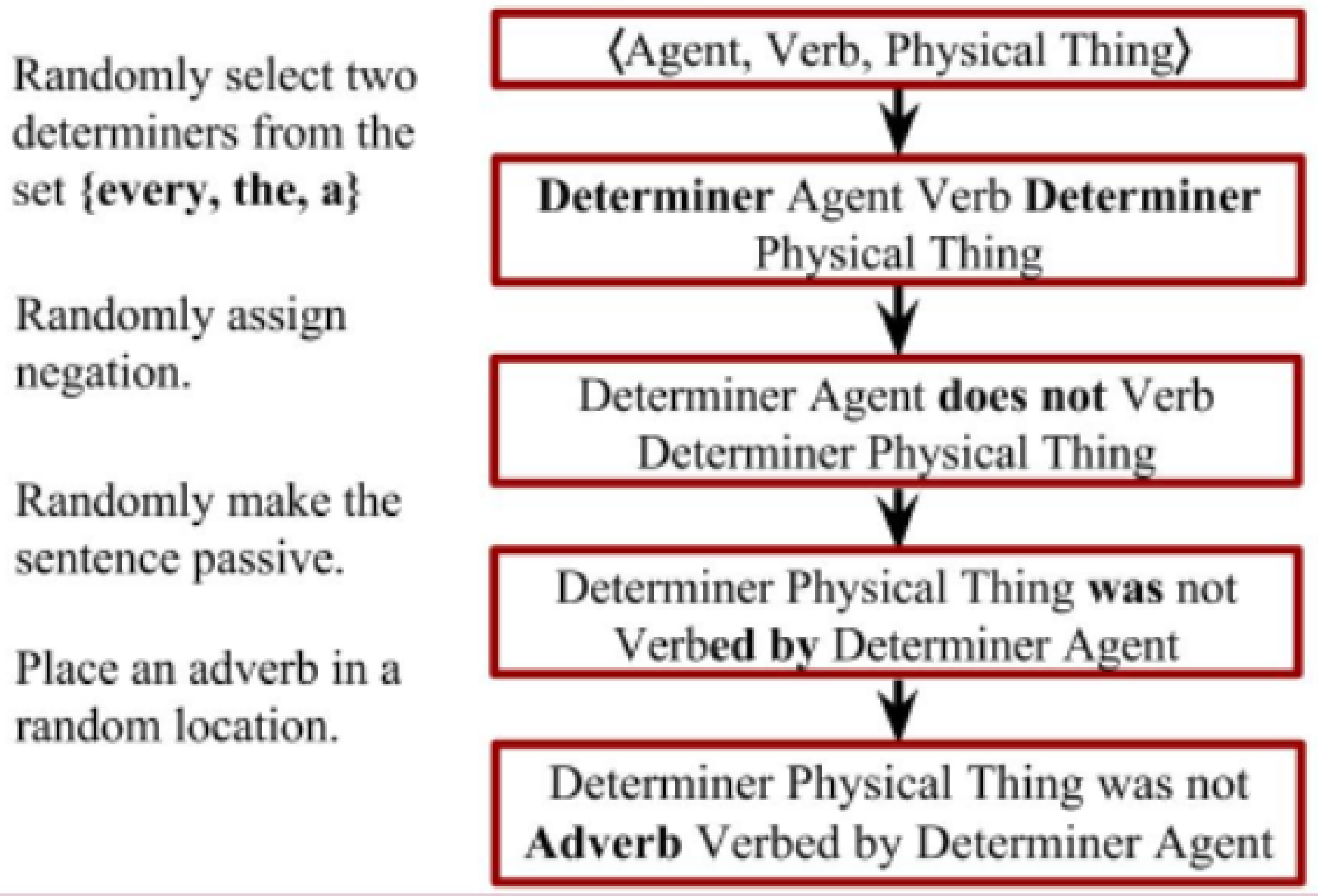
There already exist multiple NLI corpora for the evaluation of learning models. The Stanford NLI corpus is a massive corpus created by Amazon Mechanical Turks with over 500,000 labeled pairs of sentences. The SICK corpus is a automatically generated corpus with 5,000 labeled pairs of sentences. The PASCAL RTE challenges are small sets of hand generated data curated by linguists. A final corpus to mention is the recently released multi-genre NLI corpus which was also created by Amazon Mechanical Turks and has over 400,000 labeled pairs of sentences.

## The Automated and Exhaustive NLI corpus

The AENLI corpus is for the evaluation of learning models. It is automatically generated and large scale, consisting of over 1 million labeled examples. This corpus exhaustively explores a small precisely defined set of examples. Every premise and hypothesis sentence consist of clauses that contain a transitive verb, an optional adverb, a subject that is an agent, an object that is a physical thing and are optionally negated or passive. These independent clauses can either form a whole sentence or be conjoined using *and*, *or*, or *if...then*. The structure of the premise and hypothesis sentences are determined independently and randomly before being labeled. This results in every possible example that has this form has a chance of being generated.

## Generating Clauses

The authors of this paper hand generated 100 agentive nouns, 100 physical things, and 100 transitive verbs that are coherent when used with an agentive subject and physical thing direct object. We also generated 50 adverbs that coherently modify such verbs. A clause is generated from a ⟨Agent, Verb, Physical Thing⟩ triplet in the following way..



Randomly select two determiners from the set **{every, the, a}**

Randomly assign negation.

Randomly make the sentence passive.

Place an adverb in a random location.

⟨Agent, Verb, Physical Thing⟩

**Determiner** Agent Verb **Determiner** Physical Thing

Determiner Agent **does not** Verb Determiner Physical Thing

Determiner Physical Thing **was** not Verb**ed by** Determiner Agent

Determiner Physical Thing was not **Adverb** Verbed by Determiner Agent

## Generating Examples

From every ⟨Agent, Verb, Physical Thing⟩ triplet two clauses are generated independently to serve as a hypothesis and premise. The label is then automatically determined to create an example. Examples with different nouns and verbs are added as well, always labeled “permits”.

## Expanding the Data with Boolean Logic

| | | | | | |
|---|---|---|---|---|---|
| **A** | **E** | | **A** or **D** | **A** or **D** | **A** and/or **C** |
| entails | contradicts | | entails | entails | entails |
| **B** | **F** | ⇒ | If **B** then **C** | If **B** then **C** | **B** and/or **D** |
| **C** | **G** | | If **B** then **E** | If **B** then **C** | **E** and/or **G** |
| entails | contradicts | | entails | entails | contradicts |
| **D** | **H** | | If **A** then **F** | If **A** then **D** | **F** and/or **H** |

Two examples with simple sentences are used to create an example with compound sentences using the rules to the left.

## Evaluating Models

We evaluated the performance a baseline logistic regression model, a seq2seq LSTM model, and a seq2seq LSTM model with attention. Glove word vectors were used for the neural models and the features for logistic regression were BLEU scores, the length difference, word overlap, unigram/bigram indicators, and cross-unigram. Below is the attention model implemented from Rocktaschel et al. (2016), where P is the encoded premise, H is the encoded hypothesis and F is the final encoding.

$$M = \tanh(W^{y}P+W^{h}H_{F})$$
$$A = \mathrm{softmax}(w^{T}M)$$
$$P_{A}=PA^{T}$$
$$F = \tanh(W^{p}P_{A}+W^{x}H_{F})$$

| | Test Disjoint/Joint | Train |
|---|---|---|
| L. R. | 71.3/78.5% | 99.1% |
| seq2seq | 88.8/89.4% | 89.4% |
| Attention | 90.2/91.2% | 90.9% |

## Discussion

The SNLI and MultiNLI corpora were created to address perceived drawbacks of existing NLI corpora, namely that they are too small for modern models requiring large amounts of data, they include automatically generated data, and they suffer from indeterminacies of event and entity coreference that harm annotation quality. The AENLI addresses the problem of size and indeterminacies through an automatic generation process that can produce a large amount of data and follows precisely defined assumptions that remove all indeterminacies. While AENLI is automatically generated and cannot claim to have the diversity in data of SNLI, AENLI offers a precise description of exactly what is being learned. The AENLI also requires reasoning involving first order logic at the word level and boolean logic at the clause level, presenting an interesting challenge to learning models.

## Future

In the coming months, adjectives will be added to the corpus creation process. Additionally, we will experiment with new neural network models for this data set.

## References

(1) Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014b. A SICK cure for the evaluation of compositional distributional semantic models. In Proc. LREC.

(2) Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

(3) Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. CoRR abs/1704.05426.

(4) [Rocktaschel et al.2016] ˈ Tim Rocktaschel, Edward ˝ Grefenstette, Karl Moritz Hermann, Toma´s Ko ˇ cisk ˇ y,´ and Phil Blunsom. 2016. Reasoning about entailment with neural attention. In Proceedings of the 2016 ICLR, San Juan, Puerto Rico.