A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference

Adina Williams [*1], **Nikita Nangia** [†2], **Samuel R. Bowman** [‡1,2,3], Kim Deng [1]

[1]Department of Linguistics
New York University
[2]Center for Data Science
New York University
[3]Department of Computer Science
New York University

Abstract

This paper introduces the Multi-Genre Natural Language Inference (MutltiNLI) corpus, a dataset deisgned for use in the development and evaluation of machine learning models for sentence understanding. At 433k examples, this resource is one of the largest corpora avaiable fo natural language inference (a.k.a *recognizing textual entailment*), imporving upon avaible resources in both its coverage and difficulty. MultiNLI accomplishes this by offering dafa from ten distinct genrs of written and spoken English, making it possible to evaluate systems on nearly the full complexity of the language, while supplying an explicit setting for evaluating cross-genre domain adaptation. In addition, an evaluation using existing machine learning models designed for the Stanford NLI corpus shows that it represents a substantially more difficult task than does that corpus, despite the two showing similar levels of interannotator agreement.

1 Introduction

Many of the most actively studied problems in NLP, including question answering, translation, and dialog, depend in large part on natural language understanding (NLU) for success. While there has been a great deal of work that uses representation learning techniques to pursue progress on these applied NLU problems directly, in order for a representation learning model to fully succeed at one of these problmes, it must simultaneously succeed both at NLU, and at one or more additional hard marchine learning problems like structured prodiction or memory access. This makes it difficult to accurately juege the degree to which current models extract reasonable representations of language meaningn in these settings.

The task of natural language inference (NLI) is well positioned to serve as a benchmark task for research on NLU. In this task, also known as *recognizing textual entailment* (Cooper et al., 1996; Fyodorov et al., 2000; Condoravdi et al., 2003; Bos and Market, 2005; Dagan et al., 2006; MacCartney and Manning, 2009), a model is presented with a pair of sentences—-like one of those in Figure 1—-and asked to judge the relationship between their meanings by picking a label from a small set: typically ENTAILMENT, NEUTRAL, CONTRADICTION. Succeeding at NLI does not require a system to solve any difficult machine learning problems except, crucially, that of extracting effective and thorough representatios for the meanings of sentences (i.e., their lexical and compositional semantics). In particular, a model must handle phenomena like lexical entailment, quantification, coreference, tense, belief, modality, and lixical and syntactic ambiguity.

As the only large human-annotated corpus for NLI currently avaiable, the Stanford NLI Corpus (SNLI; Bowman et al., 2015) has enalbed a good deal of progress on NLU, serving as a major benchmark for machine learning work on sentence understanding and spurring work on core representation learning techniques for NLU, such as attention(Wang and Jiang, 2016; Parikh et al., 2016),

*adinawilliams@nyu.edu
†niktanangia@nyu.edu
‡bowman@nyu.edu

rectangle	triangle	line
3	4	5
1	2	3

$(a^2 + b^2 = c^2)$

2

2.1

2.1.1

memory (Munkhdalai and Yu, 2017), and the use of parse stucture (Mou et al., 2016b; Bowman et al., 2016; Chen et al., 2017). However, SNLI falls short of providing a sufficient testing ground for machine learning models in two ways.

[1]

Capacity constraint		Optimal solutions					
Cash constraint (Our model)	x_t	1	0	1	1	0	0
	y_t	20	0	77	20	100	0
	w_t	25	18	0	0	0	0
	Ed_t	30	33	41	55	45	55
	I_t	15	0	35	0	55	0
	B_t	0	330	648	1177	952	1942

First, the sentences in SNLI are derived from only a single text genre—-image captios—-and are thus limited to descriptions of concrete visual scenes, rendering the hypothesis sentences used to describe these scenes short and simple, and rendering many important phenomena—-like temporal reasoning (e.g., *yesterday*), bilief (e.g., *know*), and modality e.g., *should*—-rare enough to be irrelevant to task performance. Second, because of these issues, SNLI is not sufficiently demanding to serve as an effective benchmark for NLU, with the best current model performance falling within a few percentage points of human accuracy and limited room left for fine-grained comparisons between strong models.

This paper introduces a new challenge dataset, the Multi-Genre NLI Corpus (MultiNLI), whose chief purpose is to remedy these limitations by making it possible to run large-scale NLI evaluations that capture more of the complexity of modern English. While its size (433 pairs) and mode of collection are modeled closely on SNLI, unlike that corpus, MultiNLI represents both written and spoken speech in a wide range of styles, degrees of formality, and topics.

$$ {}^{14}_{2}\mathbf{C} \tag{1}$$

$$ {}^{b}_{a}\sum A_n \tag{2}$$

$$ {}^{227}_{90}Th+ \tag{3}$$

nlp

[1]

Met my first girlfriend that way.	FACE-TO-FACE **contradiction** CCNC	I didnt meet my first girlfriend until
8 million in relief in the form of emergency housing.	GOVERNMENT **neutral** NNNN	The 8 million dollars for emergency
8 million in relief in the form of emergency housing.	GOVERNMENT **neutral** NNNN	The 8 million dollars for emergency