A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference

Adina Williams[1]
adinawilliams@nyu.edu

Nikita Nangia[2]
nikitanangia@nyu.edu

Samuel R. Bowman[1,2,3]
bowman@nyu.edu

[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 (MultiNLI) corpus, a dataset designed 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 available for natural language inference (a.k.a. *recognizing textual entailment*), improving upon available resources in both its coverage and difficulty. MultiNLI accomplishes this by offering data from ten distinct genres 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 inter-annotator 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 problems, it must simultaneously succeed both at NLU, and at one or more additional hard machine learning problems like structured prediction or memory access. This makes it difficult to accurately judge the degree to which current models extract reasonable representations of language meaning 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 Markert, 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, and CONTRADICTION. Succeeding at NLI does not require a system to solve any difficult machine learning problems except, crucially, that of extracting effective and thorough representations 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 lexical and syntactic ambiguity.

As the only large human-annotated corpus for NLI currently available, the Stanford NLI Corpus (SNLI; Bowman et al., 2015) has enabled 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), memory (Munkhdalai and Yu, 2017), and the use of parse structure (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.

Proceedings of NAACL-HLT 2018, pages 1112–1122
New Orleans, Louisiana, June 1 - 6, 2018.

Met my first girlfriend that way.	FACE-TO-FACE **contradiction** C C N C	I didn't meet my first girlfriend until later.
8 million in relief in the form of emergency housing.	GOVERNMENT **neutral** N N N N	The 8 million dollars for emergency housing was still not enough to solve the problem.
Now, as children tend their gardens, they have a new appreciation of their relationship to the land, their cultural heritage, and their community.	LETTERS **neutral** N N N N	All of the children love working in their gardens.
At 8:34, the Boston Center controller received a third transmission from American 11	9/11 **entailment** E E E E	The Boston Center controller got a third transmission from American 11.
I am a lacto-vegetarian.	SLATE **neutral** N N E N	I enjoy eating cheese too much to abstain from dairy.
someone else noticed it and i said well i guess that's true and it was somewhat melodious in other words it wasn't just you know it was really funny	TELEPHONE **contradiction** C C C C	No one noticed and it wasn't funny at all.

Table 1: Randomly chosen examples from the development set of our new corpus, shown with their genre labels, their selected gold labels, and the validation labels (abbreviated E, N, C) assigned by individual annotators.

First, the sentences in SNLI are derived from only a single text genre—image captions—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*), belief (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 (433k 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.

Our chief motivation in creating this corpus is to provide a benchmark for ambitious machine learning research on the core problems of NLU, but we are additionally interested in constructing a corpus that facilitates work on domain adaptation and cross-domain transfer learning. These techniques—which use labeled training data for a source domain, and aim to train a model that performs well on test data from a target domain with a different distribution—have resulted in gains across many tasks (Daume III and Marcu, 2006; Ben-David et al., 2007), including sequence and part-of-speech tagging (Blitzer et al., 2006; Peng and Dredze, 2017). Moreover, in application areas outside NLU, artificial neural network techniques have made it possible to train general-purpose feature extractors that, with no or minimal retraining, can extract useful features for a variety of styles of data (Krizhevsky et al., 2012; Zeiler and Fergus, 2014; Donahue et al., 2014). However, attempts to bring this kind of general purpose representation learning to NLU have seen only very limited success (see, for example, Mou et al., 2016a). Nearly all successful applications of representation learning to NLU have involved models that are trained on data closely resembling the target evaluation data in both task and style. This fact limits the usefulness of these tools for problems involving styles of language not represented in large annotated training sets.

With this in mind, we construct MultiNLI so as to make it possible to explicitly evaluate models both on the quality of their sentence representations within the training domain and on their ability to derive reasonable representations in unfamiliar domains. The corpus is derived from ten different genres of written and spoken English, which are collectively meant to approximate the full diversity of ways in which modern standard

This task will involve reading a line from a non-fiction article and writing three sentences that relate to it. The line will describe a situation or event. Using only this description and what you know about the world:

- Write one sentence that is definitely correct about the situation or event in the line.
- Write one sentence that might be correct about the situation or event in the line.
- Write one sentence that is definitely incorrect about the situation or event in the line.

Figure 1: The main text of a prompt (truncated) that was presented to our annotators. This version is used for the written non-fiction genres.

American English is used. All of the genres appear in the test and development sets, but only five are included in the training set. Models thus can be evaluated on both the *matched* test examples, which are derived from the same sources as those in the training set, and on the *mismatched* examples, which do not closely resemble any of those seen at training time.

2 The Corpus

2.1 Data Collection

The data collection methodology for MultiNLI is similar to that of SNLI: We create each sentence pair by selecting a premise sentence from a preexisting text source and asking a human annotator to compose a novel sentence to pair with it as a hypothesis. This section discusses the sources of our premise sentences, our collection method for hypotheses, and our validation (relabeling) strategy.

Premise Text Sources The MultiNLI premise sentences are derived from ten sources of freely available text which are meant to be maximally diverse and roughly represent the full range of American English. We selected nine sources from the second release of the Open American National Corpus (OANC; Fillmore et al., 1998; Macleod et al., 2000; Ide and Macleod, 2001; Ide and Suderman, 2006, downloaded 12/2016[1]), balancing the volume of source text roughly evenly across genres, and avoiding genres with content that would be too difficult for untrained annotators.

OANC data constitutes the following nine genres: transcriptions from the Charlotte Narrative and Conversation Collection of two-sided, in-person conversations that took place in the early 2000s (FACE-TO-FACE); reports, speeches, letters, and press releases from public domain government websites (GOVERNMENT); letters from the Indiana Center for Intercultural Communication of Philanthropic Fundraising Discourse written in the late 1990s–early 2000s (LETTERS); the public report from the National Commission on Terrorist Attacks Upon the United States released on July 22, 2004[2] (9/11); five non-fiction works on the textile industry and child development published by the Oxford University Press (OUP); popular culture articles from the archives of Slate Magazine (SLATE) written between 1996–2000; transcriptions from University of Pennsylvania's Linguistic Data Consortium Switchboard corpus of two-sided, telephone conversations that took place in 1990 or 1991 (TELEPHONE); travel guides published by Berlitz Publishing in the early 2000s (TRAVEL); and short posts about linguistics for non-specialists from the Verbatim archives written between 1990 and 1996 (VERBATIM).

For our tenth genre, FICTION, we compile several freely available works of contemporary fiction written between 1912 and 2010, spanning various genres, including mystery (*The Mysterious Affair at Styles*,[3] Christie, 1921; *The Secret Adversary*,[4] Christie, 1922; *Murder in the Gun Room*,[5] Piper, 1953), humor (*Password Incorrect*,[6] Name, 2008), western (*Rebel Spurs*,[7] Norton, 1962), science fiction (*Seven Swords*,[8] Shea, 2008; *Living History*,[9] Essex, 2016; *The Sky Is Falling*,[10] Del Rey, 1973; *Youth*,[11] Asimov, May 1952), and adventure (*Captain Blood*,[12] Sabatini, 1922).

We construct premise sentences from these ten source texts with minimal preprocessing; unique the sentences within genres, exclude very short

[1] http://www.anc.org/

[2] https://9-11commission.gov/

[3] gutenberg.org/files/863/863-0.txt

[4] gutenberg.org/files/1155/1155-0.txt

[5] gutenberg.org/files/17866/17866.txt

[6] http://manybooks.net/pages/namenother09password_incorrect/0.html

[7] gutenberg.org/files/20840/20840-0.txt

[8] http://mikeshea.net/stories/seven_swords.html, shared with the author's permission.

[9] manybooks.net/pages/essexbother10living_history/0.html

[10] gutenberg.org/cache/epub/18768/pg18768.txt

[11] gutenberg.org/cache/epub/31547/pg31547.txt

[12] gutenberg.org/files/1965/1965-0.txt

sentences (under eight characters), and manually remove certain types of non-narrative writing, such as mathematical formulae, bibliographic references, and lists.

Although SNLI is collected in largely the same way as MultiNLI, and is also permissively licensed, we do not include SNLI in the MultiNLI corpus distribution. SNLI can be appended and treated as an unusually large additional CAPTIONS genre, built on image captions from the Flickr30k corpus (Young et al., 2014).

Hypothesis Collection To collect a sentence pair, we present a crowdworker with a sentence from a source text and ask them to compose three novel sentences (the hypotheses): one which is necessarily true or appropriate whenever the premise is true (paired with the premise and labeled ENTAILMENT), one which is necessarily false or inappropriate whenever the premise is true (CONTRADICTION), and one where neither condition applies (NEUTRAL). This method of data collection ensures that the three classes will be represented equally in the raw corpus.

The prompts that surround each premise sentence during hypothesis collection are slightly tailored to fit the genre of that premise sentence. We pilot these prompts prior to data collection to ensure that the instructions are clear and that they yield hypothesis sentences that fit the intended meanings of the three classes. There are five unique prompts in total: one for written non-fiction genres (SLATE, OUP, GOVERNMENT, VERBATIM, TRAVEL; Figure 1), one for spoken genres (TELEPHONE, FACE-TO-FACE), one for each of the less formal written genres (FICTION, LETTERS), and a specialized one for 9/11, tailored to fit its potentially emotional content. Each prompt is accompanied by example premises and hypothesis that are specific to each genre.

Below the instructions, we present three text fields—one for each label—followed by a field for reporting issues, and a link to the frequently asked questions (FAQ) page. We provide one FAQ page per prompt. FAQs are modeled on their SNLI counterparts (supplied by the authors of that work) and include additional curated examples, answers to genre-specific questions arising from our pilot phase, and information about logistical concerns like payment.

For both hypothesis collection and validation, we present prompts to annotators using Hybrid (gethybrid.io), a crowdsoucring platform similar to the Amazon Mechanical Turk platform used for SNLI. We used this platform to hire an organized group of workers. 387 annotators contributed through this group, and at no point was any identifying information about them, including demographic information, available to the authors.

Statistic	SNLI	MultiNLI
Pairs w/ unanimous gold label	58.3%	58.2%
Individual label = gold label	89.0%	88.7%
Individual label = author's label	85.8%	85.2%
Gold label = author's label	91.2%	92.6%
Gold label ≠ author's label	6.8%	5.6%
No gold label (no 3 labels match)	2.0%	1.8%

Table 2: Key validation statistics for SNLI (copied from Bowman et al., 2015) and MultiNLI.

Validation We perform an additional round of annotation on test and development examples to ensure accurate labelling. The validation phase follows the same procedure used for SICK (Marelli et al., 2014b) and SNLI: Workers are presented with pairs of sentences and asked to supply a single label (ENTAILMENT, CONTRADICTION, NEUTRAL) for the pair. Each pair is relabeled by four workers, yielding a total of five labels per example. Validation instructions are tailored by genre, based on the main data collection prompt (Figure 1); a single FAQ, modeled after the validation FAQ from SNLI, is provided for reference. In order to encourage thoughtful labeling, we manually label one percent of the validation examples and offer a $1 bonus each time a worker selects a label that matches ours.

For each validated sentence pair, we assign a *gold label* representing a majority vote between the initial label assigned to the pair by the original annotator, and the four additional labels assigned by validation annotators. A small number of examples did not receive a three-vote consensus on any one label. These examples are included in the distributed corpus, but are marked with '–' in the gold label field, and should not be used in standard evaluations. Table 2 shows summary statistics capturing the results of validation, alongside corresponding figures for SNLI. These statistics indicate that the labels included in MultiNLI are about as reliable as those included in SNLI, despite MultiNLI's more diverse text contents.

Genre	#Examples Train	Dev.	Test	#Wds. Prem.	'S' parses Prem.	Hyp.	Agrmt.	Model Acc. ESIM	CBOW
SNLI	*550,152*	*10,000*	*10,000*	*14.1*	*74%*	*88%*	*89.0%*	*86.7%*	*80.6 %*
FICTION	77,348	2,000	2,000	14.4	94%	97%	89.4%	73.0%	67.5%
GOVERNMENT	77,350	2,000	2,000	24.4	90%	97%	87.4%	74.8%	67.5%
SLATE	77,306	2,000	2,000	21.4	94%	98%	87.1%	67.9%	60.6%
TELEPHONE	83,348	2,000	2,000	25.9	71%	97%	88.3%	72.2%	63.7%
TRAVEL	77,350	2,000	2,000	24.9	97%	98%	89.9%	73.7%	64.6%
9/11	0	2,000	2,000	20.6	98%	99%	90.1%	71.9%	63.2%
FACE-TO-FACE	0	2,000	2,000	18.1	91%	96%	89.5%	71.2%	66.3%
LETTERS	0	2,000	2,000	20.0	95%	98%	90.1%	74.7%	68.3%
OUP	0	2,000	2,000	25.7	96%	98%	88.1%	71.7%	62.8%
VERBATIM	0	2,000	2,000	28.3	93%	97%	87.3%	71.9%	62.7%
MultiNLI Overall	**392,702**	**20,000**	**20,000**	**22.3**	**91%**	**98%**	**88.7%**	**72.2%**	**64.7%**

Table 3: Key statistics for the corpus by genre. The first five genres represent the *matched* section of the development and test sets, and the remaining five represent the *mismatched* section. The first three statistics provide the number of examples in each genre. *#Wds. Prem.* is the mean token count among premise sentences. *'S' parses* is the percentage of sentences for which the Stanford Parser produced a parse rooted with an 'S' (sentence) node. *Agrmt.* is the percent of individual labels that match the gold label in validated examples. *Model Acc.* gives the test accuracy for ESIM and CBOW models (trained on either SNLI or MultiNLI), as described in Section 3.

2.2 The Resulting Corpus

Table 1 shows randomly chosen development set examples from the collected corpus. Hypotheses tend to be fluent and correctly spelled, though not all are complete sentences. Punctuation is often omitted. Hypotheses can rely heavily on knowledge about the world, and often don't correspond closely with their premises in syntactic structure.

Unlabeled test data is available on Kaggle for both matched and mismatched sets as competitions that will be open indefinitely; Evaluations on a subset of the test set have previously been conducted with different leaderboards through the RepEval 2017 Workshop (Nangia et al., 2017).

The corpus is available in two formats—tab separated text and JSON Lines (`jsonl`), following SNLI. For each example, premise and hypothesis strings, unique identifiers for the pair and prompt, and the following additional fields are specified:

- `gold_label`: label used for classification. In examples rejected during the validation process, the value of this field will be '–'.
- `sentence{1,2}_parse`: Each sentence as parsed by the Stanford PCFG Parser 3.5.2 (Klein and Manning, 2003).
- `sentence{1,2}_binary_parse`: parses in unlabeled binary-branching format.
- `label[1]`: The label assigned during the creation of the sentence pair. In rare cases this may be different from `gold_label`, if a consensus of annotators chose a different label during the validation phase.
- `label[2...5]`: The four labels assigned during validation by individual annotators to each development and test example. These fields will be empty for training examples.

The current version of the corpus is freely available at `nyu.edu/projects/bowman/multinli/` for typical machine learning uses, and may be modified and redistributed. The majority of the corpus is released under the OANC's license, which allows all content to be freely used, modified, and shared under permissive terms. The data in the FICTION section falls under several permissive licenses; *Seven Swords* is available under a Creative Commons Share-Alike 3.0 Unported License, and with the explicit permission of the author, *Living History* and *Password Incorrect* are available under Creative Commons Attribution 3.0 Unported Licenses; the remaining works of fiction are in the public domain in the United States (but may be licensed differently elsewhere).

Partition The distributed corpus comes with an explicit train/test/development split. The test and development sets contain 2,000 randomly selected examples each from each of the genres, resulting in a total of 20,000 examples per set. No premise sentence occurs in more than one set.

Train	Model	SNLI	MNLI Match.	MNLI Mis.
	Most freq.	34.3	36.5	35.6
SNLI	CBOW	80.6	-	-
	BiLSTM	81.5	-	-
	ESIM	**86.7**	-	-
MNLI	CBOW	51.5	64.8	64.5
	BiLSTM	50.8	66.9	66.9
	ESIM	60.7	**72.3**	**72.1**
MNLI+ SNLI	CBOW	74.7	65.2	64.6
	BiLSTM	74.0	67.5	67.1
	ESIM	79.7	**72.4**	**71.9**

Table 4: Test set accuracies (%) for all models; *Match.* represents test set performance on the MultiNLI genres that are also represented in the training set, *Mis.* represents test set performance on the remaining ones; *Most freq.* is a trivial 'most frequent class' baseline.

Statistics Table 3 shows some additional statistics. Premise sentences in MultiNLI tend to be longer (max 401 words, mean 22.3 words) than their hypotheses (max 70 words, mean 11.4 words), and much longer, on average, than premises in SNLI (mean 14.1 words); premises in MultiNLI also tend to be parsed as complete sentences at a much higher rate on average (91%) than their SNLI counterparts (74%). We observe that the two spoken genres differ in this—with FACE-TO-FACE showing more complete sentences (91%) than TELEPHONE (71%)—and speculate that the lack of visual feedback in a telephone setting may result in a high incidence of interrupted or otherwise incomplete sentences.

Hypothesis sentences in MultiNLI generally cannot be derived from their premise sentences using only trivial editing strategies. While 2.5% of the hypotheses in SNLI differ from their premises by deletion, only 0.9% of those in MultiNLI (170 examples total) are constructed in this way. Similarly, in SNLI, 1.6% of hypotheses differ from their premises by addition, substitution, or shuffling a single word, while in MultiNLI this only happens in 1.2% of examples. The percentage of hypothesis-premise pairs with high token overlap (>37%) was comparable between MultiNLI (30% of pairs) and SNLI (29%). These statistics suggest that MultiNLI's annotations are comparable in quality to those of SNLI.

3 Baselines

To test the difficulty of the corpus, we experiment with three neural network models. The first is a simple continuous bag of words (CBOW) model in which each sentence is represented as the sum of the embedding representations of its words. The second computes representations by averaging the states of a bidirectional LSTM RNN (BiLSTM; Hochreiter and Schmidhuber, 1997) over words. For the third, we implement and evaluate Chen et al.'s Enhanced Sequential Inference Model (ESIM), which is roughly tied for the state of the art on SNLI at the time of writing. We use the base ESIM without ensembling with a TreeLSTM (as in the 'HIM' runs in that work).

The first two models produce separate vector representations for each sentence and compute label predictions for pairs of representations. To do this, they concatenate the representations for premise and hypothesis, their difference, and their element-wise product, following Mou et al. (2016b), and pass the result to a single tanh layer followed by a three-way softmax classifier.

All models are initialized with 300D reference GloVe vectors (840B token version; Pennington et al., 2014). Out-of-vocabulary (OOV) words are initialized randomly and word embeddings are fine-tuned during training. The models use 300D hidden states, as in most prior work on SNLI. We use Dropout (Srivastava et al., 2014) for regularization. For ESIM, we use a dropout rate of 0.5, following the paper. For CBOW and BiLSTM models, we tune Dropout on the SNLI development set and find that a drop rate of 0.1 works well. We use the Adam (Kingma and Ba, 2015) optimizer with default parameters. Code is available at `github.com/nyu-mll/multiNLI/`.

We train models on SNLI, MultiNLI, and a mixture; Table 4 shows the results. In the mixed setting, we use the full MultiNLI training set and randomly select 15% of the SNLI training set at each epoch, ensuring that each available genre is seen during training with roughly equal frequency.

We also train a separate CBOW model on each individual genre to establish the degree to which simple models already allow for effective transfer across genres, using a dropout rate of 0.2. When training on SNLI, a single random sample of 15% of the original training set is used. For each genre represented in the training set, the model that performs best on it was trained on that genre; a model trained only on SNLI performs worse on every genre than comparable models trained on any genre from MultiNLI.

Models trained on a single genre from MultiNLI perform well on similar genres; for example, the model trained on TELEPHONE attains the best accuracy (63%) on FACE-TO-FACE, which was nearly one point better than it received on itself. SLATE seems to be a difficult and relatively unusual genre and performance on it is relatively poor in this setting; when averaging over runs trained on SNLI and all genres in the matched section of the training set, average performance on SLATE was only 57.5%. Sentences in SLATE cover a wide range of topics and phenomena, making it hard to do well on, but also forcing models trained on it be broadly capable; the model trained on SLATE achieves the highest accuracy of any model on 9/11 (55.6%) and VERBATIM (57.2%), and relatively high accuracy on TRAVEL (57.4%) and GOVERNMENT (58.3%). We also observe that our models perform similarly on both the matched and mismatched test sets of MultiNLI. We expect genre mismatch issues to become more conspicuous as models are developed that can better fit MultiNLI's training genres.

To evaluate the contribution of sentence length to corpus difficulty, we binned premises and hypotheses by length in 25-word increments for premises and 10-word increments for hypotheses. Using the ESIM model, our strong baseline, we find a small effect (stronger for matched than mismatched) of premise length on model accuracy: accuracy decreases slightly as premise sentences increase in length. We find no effect of hypothesis length on accuracy.

4 Discussion and Analysis

4.1 Data Collection

In data collection for NLI, different annotator decisions about the coreference between entities and events across the two sentences in a pair can lead to very different assignments of pairs to labels (de Marneffe et al., 2008; Marelli et al., 2014a; Bowman et al., 2015). Drawing an example from Bowman et al., the pair *"a boat sank in the Pacific Ocean"* and *"a boat sank in the Atlantic Ocean"* can be labeled either CONTRADICTION or NEUTRAL depending on (among other things) whether the two mentions of boats are assumed to refer to the same entity in the world. This uncertainty can present a serious problem for inter-annotator agreement, since it is not clear that it is possible to define an explicit set of rules around coreference that would be easily intelligible to an untrained annotator (or any non-expert).

Bowman et al. attempt to avoid this problem by using an annotation prompt that is highly dependent on the concreteness of image descriptions; but, as we engage with the much more abstract writing that is found in, for example, government documents, there is no reason to assume *a priori* that any similar prompt and annotation strategy can work. We are surprised to find that this is not a major issue. Through a relatively straightforward trial-and-error piloting phase, followed by discussion with our annotators, we manage to design prompts for abstract genres that yield high inter-annotator agreement scores nearly identical to those of SNLI (see Table 2). These high scores suggest that our annotators agreed on a single task definition, and were able to apply it consistently across genres.

4.2 Overall Difficulty

As expected, both the increase in the diversity of linguistic phenomena in MultiNLI and its longer average sentence length conspire to make MultiNLI dramatically more difficult than SNLI. Our three baseline models perform better on SNLI than MultiNLI by about 15% when trained on the respective datasets. All three models achieve accuracy above 80% on the SNLI test set when trained only on SNLI. However, when trained on MultiNLI, only ESIM surpasses 70% accuracy on MultiNLI's test sets. When we train models on MultiNLI and downsampled SNLI, we see an expected significant improvement on SNLI, but no significant change in performance on the MultiNLI test sets, suggesting including SNLI in training doesn't drive substantial improvement. These results attest to MultiNLI's difficulty, and with its relatively high inter-annotator agreement, suggest that it presents a problem with substantial headroom for future work.

4.3 Analysis by Linguistic Phenomenon

To better understand the types of language understanding skills that MultiNLI tests, we analyze the collected corpus using a set of annotation tags chosen to reflect linguistic phenomena which are known to be potentially difficult. We use two methods to assign tags to sentences. First, we use the Penn Treebank (PTB; Marcus et al., 1993) part-of-speech tag set (via the included Stanford Parser parses) to automatically isolate sentences

Tag	Dev. Freq. SNLI	Dev. Freq. MultiNLI	Diff.	Most Frequent Label Label	Most Frequent Label %	Model Acc. CBOW	Model Acc. BiLSTM	Model Acc. ESIM
Entire Corpus	100	100	0	entailment	*∼35*	*∼65*	*∼67*	*∼72*
Pronouns (PTB)	34	68	34	entailment	34	66	68	73
Quantifiers	33	63	30	contradiction	36	66	68	73
Modals (PTB)	<1	28	28	entailment	35	65	67	72
Negation (PTB)	5	31	26	contradiction	**48**	67	70	75
WH terms (PTB)	5	30	25	entailment	35	64	65	72
Belief Verbs	<1	19	18	entailment	34	64	67	71
Time Terms	19	36	17	neutral	35	64	66	71
Discourse Mark.	<1	14	14	neutral	34	62	64	70
Presup. Triggers	8	22	14	neutral	34	65	67	73
Compr./Supr.(PTB)	3	17	14	neutral	39	61	63	69
Conditionals	4	15	11	neutral	35	65	68	73
Tense Match (PTB)	62	69	7	entailment	37	67	68	73
Interjections (PTB)	<1	5	5	entailment	36	67	70	75
>20 words	<1	5	5	entailment	**42**	65	67	76

Table 5: Dev. Freq. is the percentage of dev. set examples that include each phenomenon, ordered by greatest difference in frequency of occurrence (Diff.) between MultiNLI and SNLI. Most Frequent Label specifies which label is the most frequent for each tag in the MultiNLI dev. set, and % is its incidence. Model Acc. is the dev. set accuracy (%) by annotation tag for each baseline model (trained on MultiNLI only). (PTB) marks a tag as derived from Penn Treebank-style parser output tags (Marcus et al., 1993).

containing a range of easily-identified phenomena like comparatives. Second, we isolate sentences that contain hand-chosen key words indicative of additional interesting phenomena.

The hand-chosen tag set covers the following phenomena: QUANTIFIERS contains single words with quantificational force (see, for example, Heim and Kratzer, 1998; Szabolcsi, 2010, e.g., *many, all, few, some*); BELIEF VERBS contains sentence-embedding verbs denoting mental states (e.g., *know, believe, think*), including irregular past tense forms; TIME TERMS contains single words with abstract temporal interpretation, (e.g., *then, today*) and month names and days of the week; DISCOURSE MARKERS contains words that facilitate discourse coherence (e.g., *yet, however, but, thus, despite*); PRESUPPOSITION TRIGGERS contains words with lexical presuppositions (Stalnaker, 1974; Schlenker, 2016, e.g., *again, too, anymore*[13]); CONDITIONALS contains the word *if*. Table 5 presents the frequency of the tags in SNLI and MultiNLI, and model accuracy on MultiNLI (trained only on MultiNLI).

The incidence of tags varies by genre; the percentage of sentence pairs containing a particular annotation tag differs by a maximum over 30% across genres. Sentence pairs containing pronouns are predictably common for all genres, with 93% of Government and Face-to-face pairs including at least one. The Telephone genre has the highest percentage of sentence pairs containing one occurrence of negation, WH-words, *belief*-verbs and time terms, Verbatim has the highest percentage of pairs containing quantifiers and conversational pivots, and Letters has the highest percentage of pairs that contain one or more modals. Pairs containing comparatives and/or superlatives, which is the tag that our baseline models perform worst on, are most common in the Oxford University Press genre. Based on this, we conclude that the genres are sufficiently different, because they are not uniform with respect to the percentages of sentence pairs that contain each of the annotation tags.

The distributions of labels within each tagged subset of the corpus roughly mirrors the balanced overall distribution. The most frequent class overall (in this case, ENTAILMENT) occurs with a frequency of roughly one third (see Table 4) in most. Only two annotation tags differ from the baseline percentage of the most frequent class in the corpus by at least 5%: sentences containing negation, and sentences exceeding 20 words. Sentences that contain negation are slightly more likely than average to be labeled CONTRADICTION, reflecting a similar finding in SNLI, while long sentences are slightly more likely to be labeled ENTAILMENT.

None of the baseline models perform substantially better on any tagged set than they do on the corpus overall, with average model accuracies on sentences containing specific tags falling within

[13]Because their high frequency in the corpus, extremely common triggers like *the* were excluded from this tag.

about 3 points of overall averages. Using baseline model test accuracy overall as a metric (see Table 4), our baseline models had the most trouble on sentences containing comparatives or superlatives (losing 3-4 points each). Despite the fact that 17% of sentence pairs in the corpus contained at least one instance of comparative or superlative, our baseline models don't utilize the information present in these sentences to predict the correct label for the pair, although presence of a comparative or superlative is slightly more predictive of a NEUTRAL label.

Moreover, the baseline models perform below average on discourse markers, such as *despite* and *however*, losing roughly 2 to 3 points each. Unsurprisingly, the attention-based ESIM model performs better than the other two on sentences with greater than 20 words. Additionally, our baseline models do show slight improvements in accuracy on negation, suggesting that they may be tracking it as a predictor of CONTRADICTION.

5 Conclusion

Natural language inference makes it easy to judge the degree to which neural network models for sentence understanding capture the full meanings for natural language sentences. Existing NLI datasets like SNLI have facilitated substantial advances in modeling, but have limited headroom and coverage of the full diversity of meanings expressed in English. This paper presents a new dataset that offers dramatically greater linguistic difficulty and diversity, and also serves as a benchmark for cross-genre domain adaptation.

Our new corpus, MultiNLI, improves upon SNLI in its empirical coverage—because it includes a representative sample of text and speech from ten different genres, as opposed to just simple image captions—and its difficulty, containing a much higher percentage of sentences tagged with one or more elements from our tag set of thirteen difficult linguistic phenomena. This greater diversity is reflected in the dramatically lower baseline model performance on MultiNLI than on SNLI (see Table 5) and comparable inter-annotator agreement, suggesting that MultiNLI has a lot of headroom remaining for future work.

The MultiNLI corpus was first released in draft form in the first half of 2017, and in the time since its initial release, work by others (Conneau et al., 2017) has shown that NLI can also be an effective source task for pre-training and transfer learning in the context of sentence-to-vector models, with models trained on SNLI and MultiNLI substantially outperforming all prior models on a suite of established transfer learning benchmarks. We hope that this corpus will continue to serve for many years as a resource for the development and evaluation of methods for sentence understanding.

Acknowledgments

This work was made possible by a Google Faculty Research Award. SB also gratefully acknowledges support from Tencent Holdings and Samsung Research. We also thank George Dahl, the organizers of the RepEval 2016 and RepEval 2017 workshops, Andrew Drozdov, Angeliki Lazaridou, and our other NYU colleagues for help and advice.

References

Isaac Asimov. May 1952. *Youth*. Space Science Fiction Magazine, Republic Features Syndicate, Inc.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2007. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*. pages 137–144.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 120–128. http://www.aclweb.org/anthology/W06-1615.

Johan Bos and Katja Markert. 2005. Recognising textual entailment with logical inference. In *Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pages 628–635. http://www.aclweb.org/anthology/H05-1079.

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*. Association for Computational Linguistics, pages 632–642. https://doi.org/10.18653/v1/D15-1075.

Samuel R. Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D. Manning, and Christopher Potts. 2016. A fast unified model for parsing and sentence understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers). Association for Computational Linguistics, pages 1466–1477. https://doi.org/10.18653/v1/P16-1139.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1657–1668. https://doi.org/10.18653/v1/P17-1152.

Agatha Christie. 1921. *Mysterious Affair at Styles*. The Bodley Head.

Agatha Christie. 1922. *The Secret Adversary*. Dodd, Mead.

Cleo Condoravdi, Dick Crouch, Valeria de Paiva, Reinhard Stolle, and Daniel G. Bobrow. 2003. Entailment, intensionality and text understanding. In *Proceedings of the Human Language Technology-North American Association for Computational Linguistics 2003 Workshop on Text Meaning*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364* .

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Steve Pulman Ted Briscoe Holger Maier Poesio, and Karsten Konrad. 1996. Using the framework. Technical report, Technical Report LRE 62-051 D-16, The FraCaS Consortium.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. Evaluating predictive uncertainty, visual object classification, and recognising textual entailment*, Springer, pages 177–190.

Hal Daume III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research* 26:101–126. https://doi.org/doi:10.1613/jair.1872.

Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding contradictions in text. In *Proceedings of Association for Computational Linguistics-08: Human Language Technology*. Association for Computational Linguistics, pages 1039–1047. http://www.aclweb.org/anthology/P08-1118.

Lester Del Rey. 1973. *The Sky Is Falling*. Ace Books.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. DeCAF: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Ben Essex. 2016. *Living History*. CreateSpace Independent Publishing Platform.

Charles Fillmore, Nancy Ide, Daniel Jurafsky, and Catherine Macleod. 1998. An American National Corpus: A proposal. In *Proceedings of the First Annual Conference on Language Resources and Evaluation*. pages 965–969.

Yaroslav Fyodorov, Yoad Winter, and Nissim Francez. 2000. A natural logic inference system. In *Proceedings of the 2nd Workshop on Inference in Computational Semantics*.

Irene Heim and Angelika Kratzer. 1998. *Semantics in generative grammar*. Blackwell Publishers.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Nancy Ide and Catherine Macleod. 2001. The American National Corpus: A standardized resource of American English. In *Proceedings of Corpus Linguistics*. Lancaster University Centre for Computer Corpus Research on Language, volume 3, pages 1–7.

Nancy Ide and Keith Suderman. 2006. Integrating linguistic resources: The national corpus model. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA). http://www.aclweb.org/anthology/L06-1138.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proc. ACL*. https://doi.org/10.3115/1075096.1075150.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105.

Bill MacCartney and Christopher D Manning. 2009. An extended model of natural logic. In *Proceedings of the of the Eighth International Conference on Computational Semantics*. pages 140–156.

Catherine Macleod, Nancy Ide, and Ralph Grishman. 2000. The American National Corpus: A standardized resource for American English. In *Conference on Language Resources and Evaluation (LREC)*.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational linguistics* 19(2):313–330.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014a. Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Association for Computational Linguistics, pages 1–8. https://doi.org/10.3115/v1/S14-2001.

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 *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC)*.

Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016a. How transferable are neural networks in NLP applications? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pages 479–489. https://doi.org/10.18653/v1/D16-1046.

Lili Mou, Men Rui, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2016b. Natural language inference by tree-based convolution and heuristic matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 130–136. https://doi.org/10.18653/v1/P16-2022.

Tsendsuren Munkhdalai and Hong Yu. 2017. Neural semantic encoders. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, pages 397–407. http://www.aclweb.org/anthology/E17-1038.

Nick Name. 2008. *Password Incorrect*. Published online. Author also known as Piotr Kowalczyk.

Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel R Bowman. 2017. The repeval 2017 shared task: Multi-genre natural language inference with sentence representations. In *Proceedings of RepEval 2017: The Second Workshop on Evaluating Vector Space Representations for NLP*.

Andre Norton. 1962. *The Rebel Spurs*. The World Publishing Company, Cleveland & New York.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 2249–2255. https://doi.org/10.18653/v1/D16-1244.

Nanyun Peng and Mark Dredze. 2017. Multi-task domain adaptation for sequence tagging. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. pages 91–100.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pages 1532–1543. https://doi.org/10.3115/v1/D14-1162.

Henry Beam Piper. 1953. *Murder In the Gun Room*. H.B. Piper, New York.

Rafael Sabatini. 1922. *Captain Blood*. Houghton Mifflin Company.

Philippe Schlenker. 2016. *The Cambridge Handbook of Formal Semantics*, Cambridge University Press, chapter The Semantics/Pragmatics Interface, pages 664–727. https://doi.org/10.1017/CBO9781139236157.023.

Michael Shea. 2008. *Seven Swords*. Published online.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research (JMLR)* 15:1929–1958.

Robert Stalnaker. 1974. *Semantics and Philosophy*, New York University Press, chapter Pragmatic Presupposition, pages 329–355.

Anna Szabolcsi. 2010. *Quantification*. Cambridge University Press.

Shuohang Wang and Jing Jiang. 2016. Learning natural language inference with LSTM. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 1442–1451. https://doi.org/10.18653/v1/N16-1170.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association of Computational Linguistics* 2:67–78. http://www.aclweb.org/anthology/Q14-1006.

Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*. pages 818–833.