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]
bowman@nyu.edu

[1]Department of Linguistics
New York University

[2]Center for Data 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. In addition to being one of the largest corpora available for the task of NLI, at 433k examples, this corpus improves upon available resources in its coverage; it offers data from ten distinct genres of written and spoken English—making it possible to evaluate systems on nearly the full complexity of the language—and it offers an explicit setting for the evaluation of cross-genre domain adaptation.

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 make 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 at both 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 uniquely well-positioned to serve as a benchmark task for research on NLU. In this task, also known as *recognizing textual entailment* (RTE; 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 requires a model to fully capture sentence meaning (i.e., lexical and compositional semantics) by handling complex linguistic phenomena like lexical entailment, quantification, coreference, tense, belief, modality (reasoning about *should* and *must*, for example), and lexical and syntactic ambiguity.

Because this task of natural language inference is so simply defined and yet so demanding, we argue that any model or learning technique that is capable of high-quality NLU at the sentence level can be made to succeed at NLI with minimal additional effort, and that any model or learning technique that lacks this ability is destined to fail on at least some typical examples of NLI. However, a model's success depends heavily on the corpus it was trained on, so an NLI corpus should provide a broad, challenging sample of natural language to allow the model to extract reasonable representations of sentence meaning.

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 standard benchmark for sentence understanding and spurring work on core representation learning techniques for NLU such as attention (Wang and Jiang, 2016), memory (Munkhdalai and Yu, 2017), and the use of parse structure (Mou et al., 2016b;

Met my first girlfriend that way.	FACE-TO-FACE **contradiction** C C N C	I didn't meet my first girlfriend until later.
He turned and saw Jon sleeping in his half-tent.	FICTION **entailment** N E N N	He saw Jon was asleep.
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.
In contrast, suppliers that have continued to innovate and expand their use of the four practices, as well as other activities described in previous chapters, keep outperforming the industry as a whole.	OUP **contradiction** C C C C	The suppliers that continued to innovate in their use of the four practices consistently underperformed in the industry.
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.
For more than 26 centuries it has witnessed countless declines, falls, and rebirths, and today continues to resist the assaults of brutal modernity in its time-locked, color-rich historical center.	TRAVEL **entailment** E E E E	It has been around for more than 26 centuries.
If you need this book, it is probably too late unless you are about to take an SAT or GRE.	VERBATIM **contradiction** C C C N	It's never too late, unless you're about to take a test.

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

Bowman et al., 2016).

However, it falls short of the promise laid out above. The sentences in SNLI are derived from only a single text genre—image captions—and thus are limited to descriptions of concrete visual scenes, rendering the sentences short and simple, and making the handling of many key phenomena like temporal reasoning, belief, and modality irrelevant to task performance. Because of these two factors, SNLI is not sufficiently demanding to serve as an effective benchmark for NLU, with the best current model performance (Chen et al., 2017) falling within a few percentage points of human accuracy, and limited room left for fine-grained comparisons between 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 the full complexity of English. Its size (433k pairs) and mode of collection are modeled closely on SNLI, but unlike that corpus, MultiNLI represents both written and spoken speech, and a range of styles, degrees of formality, and topics.

Our chief motivation in creating this corpus is to provide a new benchmark for ambitious machine learning research on the core problems of NLU, but we are additionally interested in constructing the corpus in a way that facilitates work on domain adaptation and cross-domain transfer learning. In many 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 successes (see, for example, Mou et al., 2016a). Nearly all successful applications of representation learning to problems in NLU have involved models that are trained on data that closely resembles the target evaluation data, both in 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 explicitly evaluate models both on the quality of their representations of text in any of the training genres, and on their ability to derive reasonable representations outside those genres. In particular, the MultiNLI corpus consists of sentences derived from ten different sources of text, reflecting ten different genres of written and spoken English. All of the sources are present in 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 seen at training time.

We claim that MultiNLI is a much more challenging corpus than SNLI, and show it has a measurably more diverse range of linguistic phenomena and greater headroom for improvements in model performance. We illustrate the corpus's difficulty by tagging fifteen phenomena that are significantly more common in MultiNLI than in SNLI, and consequently analyzing model behavior on these phenomena. Despite its difficulty, we show that it has similar inter-annotator agreement to SNLI, which suggests it contains a reasonable but difficult sample of standard English.

In the remainder of this paper, we introduce the methods and data sources that we use to collect the corpus, introduce and analyze a number of machine learning models that are meant to provide baselines for further work on the corpus, present statistics over the resulting data, and analyze the range of linguistic phenomena present in the corpus.

2 The Corpus

2.1 Data Collection

The basic collection methodology for MultiNLI is similar to that of SNLI; sentence pairs are created 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 data collection method for hypothesis sentences, 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 cover a maximally broad range of genres of American English. Nine of these come from the second release of the Open American National Corpus (OANC; Fillmore et al., 1998; Macleod et al., 2000; Ide and Suderman, 2006, downloaded 12/2016):

- FACE-TO-FACE: Transcriptions of two-sided conversations from the Charlotte, NC area; the Charlotte Narrative and Conversation Collection (early 2000s).

- TELEPHONE: Transcriptions of two-sided conversations; University of Pennsylvania's Linguistic Data Consortium Switchboard corpus (1990–1991).

- 9/11: Report on the September 11th 2001 terrorist attacks; the National Commission on Terrorist Attacks Upon the United States (July 22, 2004).

- TRAVEL: Travel guides; Berlitz Publishing (early 2000s).

- LETTERS: Fundraising letters from non-profit organizations; The Indiana Center for Intercultural Communication of Philanthropic Fundraising Discourse (late 1990s–early 2000s).

- OUP: Five non-fiction works on the textile industry and child development; Oxford University Press.

- SLATE: Popular culture articles; the Slate Magazine archive (1996–2000).

- VERBATIM: Articles containing short posts about linguistics for non-specialists; the Verbatim archives (1990 and 1996).

- GOVERNMENT: Reports, speeches, letters, and press releases; public domain government websites.

For our tenth genre, FICTION, we compile several freely available works of contemporary American or British fiction written between 1912 and 2010. This section of our source corpus consists of eight modern works of short to mid-length fiction spanning crime, mystery, humor, western, adventure, science fiction, and fantasy. The authors of these works include Isaac Asimov, Agatha Christie, Ben Essex (Elliott Gesswell), Nick Name (Piotr Kowalczyk), Andre Norton, Lester del Ray, and Mike Shea.

We construct premise sentences from these source texts with minimal preprocessing; we select only unique sentences within genres, exclude very short sentences (under eight characters), and manually remove certain types of non-narrative writing, such as mathematical formulae, bibliographic references, and lists.

Despite the fact that SNLI is collected in largely the same way as MultiNLI, no SNLI examples are included in the distributed MultiNLI corpus. SNLI consists only of sentences derived from image captions from the Flickr30k corpus (Young et al., 2014), and thus can be treated as a large additional CAPTIONS genre.

Hypothesis Collection To collect a sentence pair using this method, we present a crowdworker with a sentence from a source text (the *premise*) and ask them to compose three novel sentences (the *hypotheses*): one which is necessarily true or appropriate in the same situations as the premise (to be paired with the premise under the label *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 each class will be represented equally in the raw corpus.

We tailor the prompts that surround each premise sentence during hypothesis collection 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

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. For the line "*The cottages near the shoreline, styled like plantation homes with large covered porches, are luxurious within; some come with private hot tubs,*" you could write "*The shoreline has plantation style homes near it, which are luxurious and often have covered porches or hot tubs.*"

- Write one sentence that might be correct about the situation or event in the line. For the line "*Government Executive magazine annually presents Government Technology Leadership Awards to recognize federal agencies and state governments for their excellent performance with information technology programs,*" you could write "*In addition to their annual Government Technology Leadership Award, Government Executive magazine also presents a cash prize for best dressed agent from a federal agency.*"

- Write one sentence that is definitely incorrect about the situation or event in the line. For the line "*Yes, he's still under arrest, which is why USAT's front-page reefer headline British Court Frees Chile's Pinochet is a bit off,*" you could write "*The headline 'British Court Free's Chile's Pinochet' is correct, since the man is freely roaming the streets.*"

The line may contain obscenities, opinions, or emotional stories about everyday life. The opinions expressed in these are not necessarily shared by the researchers or their academic affiliations. Even if you know outside information about individuals, objects, or events in the line, try to rely only the prompt itself to write sentences based on general beliefs or assumptions about what happens in the world. The New York University Center for Data Science is collecting your answers for use in research on computer understanding of English. Thank you for your help! If you have more questions, please consult our FAQ.

Figure 1: Example prompt shown to annotators on Hybrid. This prompt was used for the written non-fiction genres.

meanings of the three classes. Each prompt provides examples for the three requested hypothesis sentences (corresponding to the three labels) that are specific to the relevant genre. 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, tai-

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.

Figure 2: Distribution of sentence lengths (in words) for premise (green, right) and (purple, left) hypothesis sentences.

lored to fit its potentially emotional content.

Below the instructions, we present three text fields—one for each of *entailment*, *contradiction*, and *neutral*—followed by a fourth field for reporting issues, and a link to the frequently asked questions (FAQ) page. We provide one FAQ page tailored to each prompt. FAQs are modeled on their SNLI counterparts and include additional curated examples, answers to genre-specific questions that arose in our pilot phase (e.g., annotators need not introduce *um* into the hypotheses for spoken genres, despite their presence in the premise), and information about logistical concerns like payment.

For both hypothesis collection and validation, we present prompts to a selected pool of workers using the Hybrid crowdsourcing platform at `gethybrid.io/`. This differs from the collection of SNLI, which was done over Amazon Mechanical Turk. 387 workers contributed.

Validation We also perform an additional round of annotation on our test and development examples to ensure that their labels are accurate. The validation phase follows the procedure used in SICK (Marelli et al., 2014b) and SNLI (Bowman et al., 2015); 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. Instructions are very similar to those given in Figure 1, and a single FAQ, modeled after the validation FAQ from SNLI, is provided for reference. In order to encourage thoughtful labeling decisions, 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 and the four additional labels assigned by validation annota-
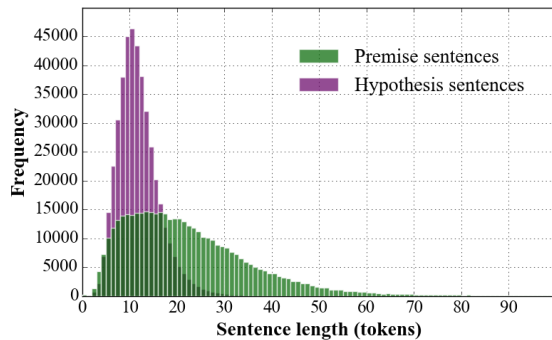
tors. A small number of examples did not receive a three-vote consensus on any one label. These examples are included in the distributed corpus, but are marked with the gold label '-', and should not be used in standard evaluations. Table 2 shows summary statistics capturing the results of validation, alongside corresponding figures for SNLI.

2.2 The Resulting Corpus

Table 1 shows one randomly chosen validated example from the development set of each of the genres along with its assigned label. Hypothesis sentences tend to rely heavily on knowledge about the world, and tend not to correspond closely with their premises in sentence structure. These syntactic differences suggest that alignment-based NLI models (like, for example, that of MacCartney, 2009) are unlikely to succeed. Hypotheses tend to be fluent and correctly spelled, consisting of full sentences, fragments, noun phrases, bare prepositional phrases, and verb phrases. Hypothesis-internal punctuation is often omitted.

The current version of the corpus is available at `http://www.nyu.edu/projects/bowman/multinli/`. The corpus is freely available 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

Genre	#Examples			#Wds. Prem.	'S' parses		Agrmt.	Model Acc.	
	Train	Dev.	Test		Prem.	Hyp.		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, broken down by genre. The first five genres represent the *matched* section of the development and test sets, and the remaining five represent the *mismatched* section. *#Examples* shows 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 below.

Unported Licenses, and the remaining works of fiction are in the public domain in the United States (but may be licensed differently elsewhere).

The corpus is available in two formats, tab separated text and JSON Lines (`jsonl`), following SNLI. Both formats have the following fields for each example:

- `gold_label`: The label to be used for classification. In examples which are rejected during the validation process, the value of this field will be '-'. These should be ignored in typical evaluations.

- `sentence1`: The premise sentence for the pair, extracted from one of the sources described above.

- `sentence2`: The hypothesis sentence for the pair, composed by a crowd worker as a companion sentence for `sentence1`.

- `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`: The above parses in unlabeled binary-branching format.

- `promptID`: A unique identifier for each premise sentence. Note that a single premise will typically appear in three different sentence pairs, each with a different hypothesis and label.

- `pairID`: A unique identifier for each example.

- `genre`: The genre of the source text from which `sentence1` was drawn.

- `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 during the validation phase chose a different label.

- `label[2...5]`: The four labels assigned by individual annotators to each development and test example during validation. For training examples, these fields will not be filled.

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

Statistics Premise sentences in MultiNLI tend to be longer than their hypothesis sentences (see Figure 2), and much longer than premise sentences in SNLI. The longest premise was 401 words and the

Training Set	Model	SNLI	MultiNLI Matched	MultiNLI Mismatched
	Most frequent class	34.3	36.5	35.6
SNLI	CBOW	80.6	-	-
	BiLSTM	81.5	-	-
	ESIM	**86.7**	-	-
MultiNLI	CBOW	51.5	64.8	64.5
	BiLSTM	50.8	66.9	66.9
	ESIM	60.7	**72.3**	**72.1**
MultiNLI + SNLI	CBOW	74.7	66.2	64.6
	BiLSTM	74.0	67.5	67.1
	ESIM	79.7	**72.4**	**71.9**

Table 4: Test set accuracies (%) for all models trained in each of three data regimes. *Matched* represents test set performance on the MultiNLI genres that are also represented in the training set, and *mismatched* represents test set performance on the remaining genres.

longest hypothesis 70 words. Table 3 contains additional statistics. We observe that the two spoken genres differ greatly from one another in apparent structure; the parser assigns far more 'S'-labeled root nodes (indicating full sentences) to premises in FACE-TO-FACE than to those in TELEPHONE. We speculate that there might be fewer full sentences in the TELEPHONE genre because conversational turn-taking is harder without visual contact, resulting in a high incidence of simultaneous speech and making transcription difficult.

In addition, we investigate what percentage of hypothesis sentences differ from premises simply by word deletion or insertion, and how many contain a high degree of word overlap. Taking SNLI as a guideline, we find that annotators delete words only in 2% of hypotheses, utilize simple insertion in only 1% of hypotheses, and write sentences with substantial token overlap between premise and hypothesis in approximately 29% of sentences. We find MultiNLI to be comparable, with only <1% of its hypotheses (170 examples total) differ from premises only by deletion, <1% of examples (224 examples total) differ from premises simply by the addition, substitution, or movement of a single word, and 30% of hypothesis sentences (4914 examples total) have >37% token overlap with premise sentences.

We were afraid annotators might use trivial strategies to make their task easier—e.g., merely adding a negative marker to hypotheses create contradictions, replacing adjectives or adverbs with synonyms or antonyms, etc.—but the measures presented above provide evidence that we succeeded in preventing them. These results suggest that our instructions and general hypothesis collection procedure doesn't encourage verbatim copying (or copying with slight modifications), and instead encourages annotators to write hypotheses using varying sentence structures.

3 Baselines

To test the difficulty of the NLI task on this corpus, we experiment with three neural network models that are intended to be maximally comparable with strong published work on SNLI.

The first two models are built to produce a single vector representing each sentence and compute label predictions based on the two resulting vectors. To do this, they concatenate the two representations, 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. The first such model 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 uses the average of the states of a bidirectional LSTM RNN (BiLSTM; Hochreiter and Schmidhuber, 1997) over the words to compute representations.

In addition to these two baselines, we also implement and evaluate Chen et al.'s (2017) Enhanced Sequential Inference Model (ESIM), which represents the state of the art on SNLI at the time of writing. We use the base ESIM without ensembling with a TreeLSTM (as is done in one published variant of the model).

All three models are initialized with 300D reference GloVe vectors (840B token version; Pennington et al., 2014). Out-of-vocabulary (OOV) words are initialized randomly, and all word embeddings are fine-tuned during training. The mod-

Train Genre	SNLI	FICT.	TRAVEL	GOV.	TEL.	SLATE	FACE.	OUP	LETT.	9/11	VER.
SNLI	**73.9**	47.1	47.0	46.2	48.1	44.6	47.3	44.4	46.2	46.4	45.9
FICTION	51.4	**65.2**	55.1	54.4	58.3	54.2	58.8	53.0	57.1	55.5	54.3
TRAVEL	46.7	54.3	**63.4**	56.5	54.6	51.7	56.5	55.5	62.5	55.0	54.9
GOVERNMENT	44.8	55.4	55.4	**65.0**	56.7	55.2	58.9	**60.0**	63.8	55.1	56.9
TELEPHONE	47.8	59.3	56.3	55.9	**62.2**	53.5	**63.0**	54.9	59.3	53.0	55.3
SLATE	47.4	58.4	57.4	58.3	56.9	**56.9**	58.5	55.9	59.9	**55.6**	**57.2**

Table 5: Test set accuracies across all genres for CBOW models trained on the individual training genres. As expected, the best performance for each genre is obtained by the model trained on that genre.

els use 300D hidden states, as in most prior work on SNLI. We use Dropout (Srivastava et al., 2014) for regularization. For ESIM, we follow the paper and use a dropout rate of 0.5. For the 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 the default parameters.

We train models on SNLI, on MultiNLI, and on a mixture of both corpora. In the mixed setting, we use the full MultiNLI training set but downsample SNLI by randomly selecting 15% of the SNLI training set at each epoch. This ensures that each available genre has roughly equal frequency during training. Table 4 shows the results.

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. When training on SNLI, a single random sample of 15% of the original training set is used. Table 5 shows these results.[1]

4 Discussion

Data Collection In data collection for NLI, different annotator decisions about the coreference between entities and events across the pair can lead to very different assignments of labels (de Marneffe et al., 2008; Marelli et al., 2014a; Bowman et al., 2015). Drawing an example from Bowman et al. (2015), 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 intelligi-

---

[1] For the CBOW models trained on individual genres, we use a dropout rate of 0.2.

ble to an untrained annotator (or any non-expert). Bowman et al. (2015) attempt to avoid this problem by using an annotation prompt that is highly dependent on the concreteness of image descriptions, but there is no reason to suspect *a priori* that any similar prompt and annotation strategy can work for the much more abstract writing that is found in, for example, government documents. We find that this is not a major issue in MultiNLI. Through relatively straightforward trial-and-error, extensive piloting by genre, and in-depth discussion with our annotators, we are able to design prompts that yield inter-annotator agreement scores nearly identical to those of SNLI (see Table 2). This suggests that the task is probably being conceptualized in a fairly uniform way across annotators and across genres.

Corpus Difficulty Our three baseline models perform better on SNLI than on MultiNLI by about fifteen percentage points when trained on the respective datasets, suggesting MultiNLI has a good deal of headroom remaining. When the models are trained only on SNLI, all three achieve accuracy above 80% on the SNLI test set. However, when trained on MultiNLI, only ESIM surpasses 70% accuracy on either of MultiNLI's test sets. When the models are trained on MultiNLI and downsampled SNLI, their performance improves significantly on SNLI as expected, but performance on the MultiNLI test sets does not significantly change. Between MultiNLI's difficulty and its relatively high inter-annotator agreement, it presents a problem with substantial headroom for future work.

Linguistic Phenomena As expected, the increase in the diversity of linguistic phenomena in MultiNLI and its longer average sentence length conspire to make it dramatically more difficult to model than SNLI. To get an idea of the sorts of linguistically difficult phenomena that MultiNLI

Tag	SNLI	Dev. Freq. MultiNLI	Diff.	Most Freq. Label Label	%	Model Acc. (MultiNLI) CBOW	BiLSTM	ESIM
Overall	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' Words (PTB)	5	**30**	25	*entailment*	35	64	65	72
Belief Verbs	<1	**19**	18	*entailment*	34	64	67	71
Time Terms	19	**36**	17	*eutral*	35	64	66	71
Conversational Pivots	<1	**14**	14	*neutral*	34	62	64	70
Presupposition Triggers	8	**22**	14	*neutral*	34	65	67	73
Comparatives/Superlatives (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
Existentials (PTB)	5	**8**	3	*contradiction*	36	**69**	**72**	**77**

Table 6: Dev. Freq. is the percentage of each development set that contains each tag, ordered by difference in incidence between the two corpora. Most Freq. Label specifies which label is the most frequent for each tag in the MultiNLI Dev set and its incidence among the tagged examples. The incidence figure is italicized when it differs from the incidence of the most frequent label in the overall corpus by over 5%. Model Acc. is the development set accuracy (%) by annotation tag for each model trained on MultiNLI. The accuracy for the tag on which each model does best is bolded, and the accuracy for the tag on which it does worst is shown in parentheses. (PTB) marks that a tag is derived automatically from Penn Treebank tags (Marcus et al., 1993).

provides, we programmatically assign a set of annotation tags to each example in the development set. These tags allow us to determine how common these linguistically difficult phenomena are in the dataset, and how well each model performs on these difficult examples.

We use two methods to tag sentences based on the linguistic phenomena they contain. First, we use the existing Penn Treebank (PTB Marcus et al., 1993) tag set to automatically isolate sentences containing difficult linguistic phenomena (e.g., pronouns, negation, existentials, modals, comparatives and superlatives, etc.). Second, we search the corpus for hand-chosen key words indicative of interesting linguistic phenomena; descriptions of these are below:

- QUANTIFIERS: Single words with quantificational force (see Heim and Kratzer 1998; Szabolcsi 2010), e.g., *many, all, few, some*.

- BELIEF VERBS: Sentence-embedding verbs that denote mental states, e.g., *know, believe, think, contemplate* etc., including irregular past tense forms (where applicable).

- TIME TERMS: Single words that are related to abstract temporal interpretation, e.g., *then,*

today, as well as month names and days of the week e.g., *January, Wednesday*.

- CONVERSATIONAL PIVOTS: Words facilitating discourse cohesion, e.g., *yet, but, thus, despite, however*.

- PRESUPPOSITION TRIGGERS: Single words that trigger *presuppositions*—an assumption(s) that must be accepted in order for certain words to be appropriately used in a certain context (see Stalnaker 1974; Schlenker 2016)—e.g. *again, too, anymore*.[2]

- CONDITIONALS: The word *if*.

A breakdown of these automatic tags in SNLI and MultiNLI is given in Table 6, along with the model accuracy on sentences with these tags on MultiNLI (for models trained on MultiNLI alone). As expected, the state-of-the-art ESIM baseline model performs better than the other two baselines on each annotation tag; the BiLSTM performs better than CBOW on each annotation tag as well.

None of the annotation tags correlate reliably enough with specific labels to allow the model to

---

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

TAG	Matched						Mismatched			
	FICT.	TRAVEL	GOV.	TEL.	SLATE	FACE.	OUP	LETT.	9/11	VER.
Pronouns (PTB)	83	52	**93**	92	65	**93**	54	60	75	68
Quantifiers	51	62	61	72	64	58	69	57	60	**74**
Modals (PTB)	23	21	33	30	28	25	30	**40**	16	29
Negation (PTB)	30	23	25	**47**	33	34	26	26	30	10
'Wh' Words (PTB)	20	28	26	**42**	24	35	38	25	25	39
Belief Verbs	17	8	13	**46**	16	25	15	13	17	24
Time Terms	26	35	28	**52**	34	38	42	38	36	35
Conversational Pivots	11	12	11	21	14	13	15	14	11	**23**
Presupposition Triggers	18	24	21	22	20	15	27	18	**33**	24
Comparatives/Superl. (PTB)	10	22	18	15	16	10	**28**	11	20	21
Conditionals	9	12	18	17	15	10	19	**22**	9	18
Tense Match (PTB)	73	68	62	70	67	69	66	70	**84**	62
Interjections (PTB)	5	0	<1	**17**	1	13	<1	5	2	<1
>20 Words	3	5	7	5	4	4	**8**	**8**	**8**	5
Existentials (PTB)	6	13	**18**	15	7	6	6	6	8	11

Table 7: Percentages of sentence pairs (from the development set) that have been assigned each annotation tag, by genre. Across each annotation tag, every MultiNLI genre has a higher incidence of every tag than SNLI (see Table 6). Bolded numbers show the highest incidence of the tag.

use the phenomenon identified by a tag as its only feature for the tagged examples and still perform well. Only two annotation tags differed from the percentage of the most frequent class in the full corpus by more than 5%: pairs containing sentences of over 20 words and pairs containing negation. Long sentences are slightly more likely to be assigned the ENTAILMENT label by the models, while sentences that contain negation are slightly more likely than average to be assigned the CONTRADICTION label by the models, just as in SNLI, reflecting a slight annotator preference for using negation when writing sentences that contradict some premise.

The baseline models' accuracies improve by 2–3 percentage points on sentence pairs that contain negation, interjections, sentences of more than 20 words, and existentials. Improvement on sentence pairs with negation suggests that the models learn that the presence of negative words increases the likelihood that the model should assign the label CONTRADICTION to the pair; improvement on sentences with interjections might be due to the fact that they are often explicitly marked by an exclamation point at the end, or perhaps because they tend to be shorter and have simpler syntax.

Improvement on long sentences and sentences with existentials might not actually be due to the model learning something specific about them, but instead might be inflated due to their rarity in MultiNLI (see Table 7); even though this corpus includes a higher percentage of long sentences and existential sentences than SNLI, the development

set still contains only 1128 pairs with an extra long premise or hypothesis and only 1556 pairs with existentials in either the premise or the hypothesis. Additionally, as we expect, the ESIM baseline model performs substantially better than the other two on sentences with greater than 20 words, which might be expected for model architectures like this one, that include attention modules for better handling of long-distance dependencies.

On the other hand, models have trouble with certain difficult linguistic phenomena captured by our annotation tags. For example, accuracy drops 3–4 points across models for sentences containing comparatives or superlatives. Despite the fact that 17% of sentence pairs in the corpus contained at least one instance of comparative or superlative, the model does not pick up on the information present in these sentences to predict the correct label for the pair, although the presence of a comparative or superlative is slightly more predictive of a NEUTRAL label. Additionally, the baseline models perform below average on conversational pivots, losing roughly 2-3 points each. This could be due to the fact that conversational pivots often appear at the beginning and ends of sentences, to ease the transition from one sentence to another or refer to information present in a previous or following sentence, although they do appear in the MultiNLI corpus sentence-internally at clause breaks. Additionally, it could because these sentences tend to be hard to interpret without more context.

We note above that models trained on MultiNLI have relatively poor performance on SNLI. This

may be because SNLI consists largely of presentational structures like *"a black car starts up in front of a crowd of people"*, while MultiNLI has very few of these.

Difficulty by Genre Furthermore, since one of the main innovations of MultiNLI over SNLI is its broader coverage of different genres of language, we investigated the presence of various linguistic phenomena by genre. The TELEPHONE genre had the highest number of tagged phenomena, suggesting it was the most difficult genre. GOVERNMENT, FACE TO FACE, and TELEPHONE had particularly high percentages of pronouns. Spoken genres TELEPHONE and FACE TO FACE have more 'Wh' words and belief verbs than other genres. The LETTERS genre has more modals and conditionals than other genres, which is consistent with the purpose of the letters, which was to encourage fund-raising donation. Genres differed substantially in how many of their sentence pairs were assigned each annotation tag, suggesting that the set of genres contained in the corpus is reasonably diverse.

Cross-Genre Similarity Table 5 provides us with an indirect measure of similarity between genres. For each genre represented in the training set, the model that performs best was trained on that genre. We see some correlation in performance among the genres we'd expect to be similar. For example, for the FACE-TO-FACE genre, the model trained on TELEPHONE (the other spoken genre) attains the best accuracy. While SLATE seems to be a more difficult genre and performance on it is relatively poor, the model trained on SLATE achieves best accuracy on 9/11 and VERBATIM. The SLATE model also gets relatively high accuracy on the TRAVEL and GOVERNMENT genres, second only to the models trained on these respective genres. This may be because sentences in SLATE cover a wide range of topics, making it harder to do well on, but also forcing models trained on it be more broadly capable.

We also observe that our models perform similarly on both the matched and mismatched test sets of MultiNLI. Any difference in performance between the sets owing to train–test genre mismatch is likely to be obscured by the substantial uncontrolled variation in difficulty between genres. We expect that as models are developed that can better fit the training genres of MultiNLI, issues of genre

mismatch will become more conspicuous.

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 difficulty and diversity of linguistic phenomena, and also serves as a benchmark for the study of cross-genre domain adaptation.

Acknowledgments

References

Johan Bos and Katja Markert. 2005. Recognising textual entailment with logical inference. In *Proc. EMNLP*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proc. EMNLP*.

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 *Proc. ACL*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proc. ACL*.

Cleo Condoravdi, Dick Crouch, Valeria de Paiva, Reinhard Stolle, and Daniel G. Bobrow. 2003. Entailment, intensionality and text understanding. In *Proc. NAACL*.

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.

Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding contradictions in text. In *Proc. ACL*.

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 *Proc. ICML*.

Charles Fillmore, Nancy Ide, Daniel Jurafsky, and Catherine Macleod. 1998. An American national corpus: A proposal. In *Proc. LREC*.

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.

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

Nancy Ide and Keith Suderman. 2006. Integrating linguistic resources: The national corpus model. In *Proc. LREC*.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. ICLR*.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proc. ACL*.

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

Bill MacCartney. 2009. *Natural Language Inference*. Ph.D. thesis, Stanford University.

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*.

Catherine Macleod, Nancy Ide, and Ralph Grishman. 2000. The american national corpus: Standardized resources for american english. In *Proc. 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. SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proc. SemEval*.

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*.

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 *Proc. EMNLP*.

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 *Proc. ACL*.

Tsendsuren Munkhdalai and Hong Yu. 2017. Neural semantic encoders. In *Proc. EACL*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proc. EMNLP*.

Philippe Schlenker. 2016. *The Cambridge Handbook of Formal Semantics*, Cambridge University Press, chapter The Semantics/Pragmatics Interface.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *JMLR* 15.

Robert Stalnaker. 1974. *Semantics and Philosophy*, chapter Pragmatic Presupposition, pages 197–213.

Anna Szabolcsi. 2010. *Quantification*.

Shuohang Wang and Jing Jiang. 2016. Learning natural language inference with LSTM. In *Proc. NAACL*.

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. *TACL* 2.

Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Proc. ECCV*.