The RepEval 2017 Shared Task: Multi-Genre Natural Language Inference with Sentence Representations

Nikita Nangia[1]
nikitanangia@nyu.edu

Adina Williams[2]
adinawilliams@nyu.edu

Angeliki Lazaridou[3]
angeliki@deepmind.com

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

[1]Center for Data Science
New York University

[2]Department of Linguistics
New York University

[3]DeepMind

Abstract

This paper presents the results of the RepEval 2017 Shared Task, which evaluated neural network sentence representation learning models on the Multi-Genre Natural Language Inference corpus (MultiNLI) recently introduced by Williams et al. (2017). All of the five participating teams beat the bidirectional LSTM (BiLSTM) and continuous bag of words baselines reported in Williams et al.. The best single model used stacked BiLSTMs with residual connections to extract sentence features and reached 74.5% accuracy on the genre-matched test set. Surprisingly, the results of the competition were fairly consistent across the genre-matched and genre-mismatched test sets, and across subsets of the test data representing a variety of linguistic phenomena, suggesting that all of the submitted systems learned reasonably domain-independent representations for sentence meaning.

1 Introduction

The Second Workshop on Evaluating Vector Space Representations for NLP (RepEval 2017) features a shared task competition meant to evaluate natural language understanding models based on sentence encoders—that is, models that transform sentences into fixed-length vector representations and reason using those representations. Submitted systems are evaluated on the task of natural language inference (NLI, also known as recognizing textual entailment, or RTE) on the Multi-Genre NLI corpus (MultiNLI; Williams et al. 2017). Each example in the corpus consists of a pair of sentences, and systems must predict whether the relationship between the two sentences is *entailment*, *neutral* or *contradiction* in a balanced three-way classification setting.

We selected the task of NLI with the intent to evaluate as directly as possible the degree to which each model can extract and manipulate distributed representations of sentence meaning. In order for a system to perform well at natural language inference, it needs to handle nearly the full complexity of natural language understanding,[1] but its framing as a sentence-pair classification problem makes it suitable as an evaluation task for a broad range of models, and avoids issues of sequence generation, structured prediction, or memory access that can complicate evaluation in other settings.

The shared task includes two evaluations, a standard in-domain (*matched*) evaluation in which the training and test data are drawn from the same sources, and a cross-domain (*mismatched*) evaluation in which the training and test data differ substantially. This cross-domain evaluation tests the ability of submitted systems to learn representations of sentence meaning that capture broadly useful features.

This paper briefly introduces the task and dataset, presents the rules and results of the competition, and analyzes and compares the submitted systems. All the submitted systems are broadly

[1]Entailment notably does not require a system to ground its representations of sentence meaning to any outside representational system, for better or worse. For related discussion of entailment and natural language understanding see Chierchia and McConnell-Ginet (1991), Dagan et al. (2006), and MacCartney (2009).

Proceedings of the 2nd Workshop on Evaluating Vector-Space Representations for NLP, pages 1–10,
Copenhagen, Denmark, September 7–11, 2017.

Met my first girlfriend that way.	FACE-TO-FACE **contradiction**	I didn't meet my first girlfriend until later.
He turned and saw Jon sleeping in his half-tent.	FICTION **entailment**	He saw Jon was asleep.
8 million in relief in the form of emergency housing.	GOVERNMENT **neutral**	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**	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**	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**	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**	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**	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**	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**	It's never too late, unless you're about to take a test.

Table 1: Randomly chosen examples from each genre of the MultiNLI development set.

similar, and incorporate bidirectional recurrent neural networks as a key component. We find that all systems performed fairly well, outperforming a simple bidirectional LSTM (BiLSTM; Hochreiter and Schmidhuber, 1997) baseline. To our surprise, no system performed dramatically worse on the *mismatched* evaluation than on the *matched* evaluation, and all systems performed reasonably consistently across examples representing a range of linguistic phenomena, suggesting that all were able to produce systems for semantic representation which, while not perfect, were effective and not tightly adapted to any particular style of language or set of constructions.

2 Dataset

MultiNLI (Williams et al., 2017) consists of 393k pairs of sentences from a broad range of genres of written and spoken English, balanced across three labels. Each premise sentence (the first sentence in each pair) is derived from one of ten sources of text, which constitute the ten genre sections of the corpus. Each hypothesis sentence and pair label was composed by a crowd worker in response to a premise. MultiNLI was designed and collected in the style of the Stanford NLI Corpus (SNLI; Bowman et al. 2015), but covers a broader range of styles of text, rather than the relatively homogeneous captions used in SNLI.

Testing and development sets are available for all genres, with 2000 examples per set per genre. Only five genres have accompanying training sets. So, for the *matched* development and test sets, models are tested on examples derived from the same sources as those in the training set, while for the *mismatched* sets, the text source is not repre-

sented in the training data.

Table 1 presents example sentences from the corpus and Table 2 presents some key statistics. For a detailed discussion of the corpus, refer to Williams et al. (2017).

3 Shared Task Competition

The purpose of the shard task is to evaluate techniques for training and using sentence encoders. To this end, we require that all models create fixed-length vectors for each sentence with no explicitly-imposed internal structure. Alignment strategies like attention that pass information between the two encoders handling the two input sentences in a pair are not allowed. Memory models that represent sentences as variable-length sets or sequences of vectors are also not permitted. While systems that use methods like attention and structured memory are effective for NLI (Rocktäschel et al., 2016; Wang and Jiang, 2016; Chen et al., 2017a; Williams et al., 2017, i.a.), much of the variation across models of this kind lies in the way that they explicitly or implicitly align related sentences, rather than the way that they extract representations for sentences. As a result, we expect that focusing our evaluation on a restricted subset of models will yield conclusions that are more generally applicable to work on natural language understanding than would have been the case otherwise.

Additional Rules We provide competitors with labeled training and development sets, and unlabeled test sets for which they must submit labels. The development sets are meant to be used for hyperparameter tuning and model selection, and training on the development sets is not allowed. We place no limits on the use of outside training data and resources except that they be publicly available. We specifically encourage the use of the SNLI training set. Multiple submissions from the same team are allowed, up to a limit of two per day during the two-week evaluation period. Individual participants (i.e., PIs) are permitted to join multiple teams within reason, but only when each team reflects a fully independent engineering effort and each team has a different lead developer.

Evaluation Competitors had approximately ten weeks, starting with the release of the MultiNLI training and development sets, to develop their systems and two additional weeks—the evaluation period—to run their systems on the unlabeled test sets and submit results. The shared task evaluation was hosted through the Kaggle in Class platform using two competition pages—one each for the *matched*[2] and *mismatched*[3] sections of the corpus. The public leaderboard, which was displayed during the evaluation period, showed results on a random 25% of the test set labels, and the final results were computed by evaluating the two best systems from each competitor (chosen from the public leaderboard) on the remaining hidden 75% of the test set labels.

4 Results and Leaderboard

The competition results are shown in Table 3. All evaluated systems beat the BiLSTM baseline reported in Williams et al.. Furthermore, there is only a marginal gap between accuracy on *matched* and *mismatched* test sets for all systems.

The best performing single model is by Nie and Bansal, who achieve the best result on the *matched* competition and tie with Chen et al. in the *mismatched* competition. The Nie and Bansal model architecture uses stacked BiLSTMs with residual connections and, unlike the other high performing models, does not use within-sentence attention. The best performing system overall is an ensemble by Chen et al., which is based closely on the Enhanced Sequential Inference Model (ESIM; Chen et al., 2017a) but with attention only within each sentence, rather than between the two.

Looking toward the future, we also made available non-time-limited Kaggle in Class competition pages[4] to allow for further fair evaluations on the MultiNLI test sets. Note that since these evaluation sites report results on 100% of the test set, rather than the 75% used in the shared task, numbers reported on that site may differ slightly from those seen in the competition.

5 Model Comparison

All of the submitted systems are based on bidirectional LSTMs, but each system uses this core tool in a somewhat different way. This section sur-

[2] https://inclass.kaggle.com/c/multinli-matched-evaluation

[3] https://inclass.kaggle.com/c/multinli-mismatched-evaluation

[4] Matched: https://inclass.kaggle.com/c/multinli-matched-open-evaluation
Mismatched: https://inclass.kaggle.com/c/multinli-mismatched-open-evaluation

Genre	#Examples Train	Dev.	Test	#Wds. Prem.	'S' parses Prem.	Hyp.	Agrmt.	BiLSTM Acc.
SNLI	*550,152*	*10,000*	*10,000*	*14.1*	*74%*	*88%*	*89.0%*	*81.5%*
FICTION	77,348	2,000	2,000	14.4	94%	97%	89.4%	66.8%
GOVERNMENT	77,350	2,000	2,000	24.4	90%	97%	87.4%	68.0%
SLATE	77,306	2,000	2,000	21.4	94%	98%	87.1%	68.4%
TELEPHONE	83,348	2,000	2,000	25.9	71%	97%	88.3%	67.7%
TRAVEL	77,350	2,000	2,000	24.9	97%	98%	89.9%	66.8%
9/11	0	2,000	2,000	20.6	98%	99%	90.1%	68.5%
FACE-TO-FACE	0	2,000	2,000	18.1	91%	96%	89.5%	67.5%
LETTERS	0	2,000	2,000	20.0	95%	98%	90.1%	66.4%
OUP	0	2,000	2,000	25.7	96%	98%	88.1%	66.7%
VERBATIM	0	2,000	2,000	28.3	93%	97%	87.3%	67.2%
MultiNLI Overall	**392,702**	**20,000**	**20,000**	**22.3**	**91%**	**98%**	**88.7%**	**67.4%**

Table 2: Key statistics for the corpus broken down by genre, presented alongside figures from SNLI for comparison. 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 shown are the number of examples in each genre. *#Wds. Prem.* is the mean token count among premise sentences. *'S' parses* is the percentage of premises or hypotheses which the Stanford Parser labeled as full sentences rather than fragments. *Agrmt.* is the percent of individual annotator labels that match the assigned gold label used in evaluation. *BiLSTM Acc.* gives the test accuracy on the full test set for the BiLSTM baseline model trained on MultiNLI and SNLI.

Team Name	Authors	Matched	Mismatched	Model Details
alpha (ensemble)	Chen et al.	**74.9%**	**74.9%**	STACK, CHAR, ATTN., POOL, PRODDIFF
YixinNie-UNC-NLP	Nie and Bansal	74.5%	73.5%	STACK, POOL, PRODDIFF, SNLI
alpha	Chen et al.	73.5%	73.6%	STACK, CHAR, ATTN, POOL, PRODDIFF
Rivercorners (ensemble)	Balazs et al.	72.2%	72.8%	ATTN, POOL, PRODDIFF, SNLI
Rivercorners	Balazs et al.	72.1%	72.1%	ATTN, POOL, PRODDIFF, SNLI
LCT-MALTA	Vu et al.	70.7%	70.8%	CHAR, ENHEMB, PRODDIFF, POOL
TALP-UPC	Yang et al.	67.9%	68.2%	CHAR, ATTN, SNLI
BiLSTM baseline	Williams et al.	67.0%	67.6%	POOL, PRODDIFF, SNLI

Table 3: RepEval 2017 shared task competition results. The Model Details column lists some of the key strategies used in each system, using keywords: STACK: use of multilayer bidirectional RNNs, CHAR: character-level embeddings, ENHEMB: embeddings enhanced with auxiliary features, POOL: max or mean pooling over RNN states, ATTN: intra-sentence attention, PRODDIFF: elementwise sentence product and difference features in the final entailment classifier, SNLI: use of the SNLI training set.

veys the key differences between systems, and the Model Details column in Table 3 serves as a summary reference for these differences.

Depth Chen et al. and Nie and Bansal use three-layer bidirectional RNNs, while others only used single-layer RNNs. This likely contributes significantly to their good performance, as it is the most prominent feature shared only by these two top systems. They both use shortcut connections between recurrent layers to ease gradient flow, and Nie and Bansal find in an ablation study that using shortcut connections improves their performance by over 1% on both development sets.

Embeddings Systems vary reasonably widely in their approach to input encoding. Yang et al. and Chen et al. use a combination of GloVe embeddings (Pennington et al., 2014, not fine tuned) and character-level convolutional neural networks (Kim et al., 2016) to extract representations of words. Balazs et al. also use pre-trained GloVe embeddings without fine tuning, but report (contra Chen et al.) that an added character-level feature extractor does not improve performance.

Vu et al. use pre-trained GloVe word embeddings augmented with additional feature vectors. They create embeddings for part-of-speech (POS), character level information, and the dependency relation between a word and its parent, and con-

catenate these with the embedding for each word. They find that this supplies a small but nontrivial improvement to their development set performance, especially in the *mismatched* setting.

Nie and Bansal use the simplest strategy, initializing embeddings with GloVe vectors and fine-tuning them.

Pooling Vu et al. make a surprisingly effective change to the baseline BiLSTM model, motivated by Conneau et al.'s (2017) findings, by using max pooling rather than mean pooling when collecting the hidden states of the bidirectional LSTM for use as a sentence representation. They find that this yields an improvement of over 2.5% on both development sets.

While Vu et al. show that the choice of pooling strategy is quite important, Balazs et al. do not find a substantial effect in a similar comparison. This may be because Balazs et al.'s model also makes use of intra-sentence attention following the pooling layer, which dramatically reduces the importance of pooling.

Intra-Sentence Attention Chen et al. and Balazs et al. both use attention over the BiLSTM states of each sentence to compute a final representation for that sentence. Chen et al. in particular uses a novel *gated* attention formulation, in which the BiLSTM gate values supply the attention weights over hidden states according to

$$v_g = \sum_{i=1}^{n} \frac{\|g_i\|_2}{\sum_{j=1}^{n} \|g_j\|_2} h_i$$

where g_i is the BiLSTM input gate and h_i is the output from the BiLSTM encoder. They find that their use of gated attention helps performance somewhat relative to an unspecified baseline, though only in the *matched* setting.

Sentence Pair Classifier Every system but Yang et al.'s uses elementwise product and difference features, comparing the two sentence encodings as part of the input to the classifier MLP that predicts the final relation label. In an ablation study, Chen et al. find this to be highly important, yielding more than a 3% gain in performance on both development sets.

Data and SNLI We observe relatively little variation in the training data used in submitted systems. All systems are trained only on labeled NLI data—either the MultiNLI training set alone, or the MultiNLI and SNLI training sets combined. While Williams et al. find that the combined training set yields somewhat better results on the MultiNLI test set, Chen et al. nonetheless reaches state-of-the-art performance without using it.

Interim Discussion We were particularly struck by the effectiveness of the max pooling strategy as a simple and highly effective improvement to the baseline BiLSTM sentence encoder. Less surprisingly, depth and intra-sentence attention appear to be broadly effective, and product and difference features appear to be valuable when using sentence encoders for the task of NLI. The results surrounding embeddings and input encoding were less clear, though Nie and Bansal's use of pre-trained GloVe embeddings with fine tuning appears to be a simple and effective approach.

6 Error Analysis

In the interest of better understanding both the corpus and the submitted models, we annotate a 1,000-sample subset of the development set. We also provide a set of probe sentences and ask participating teams to submit vectors for all sentences in the probe set and test set. This section surveys our methods findings.

6.1 Annotations

The annotated subset of the development set was released to competitors during the model development period, and consists of 1,000 examples each tagged with zero or more of the following labels. Labels were assigned manually except where clear keyword-spotting techniques sufficed.

- CONDITIONAL: Whether either sentence contains a conditional.
  Example: **P:** *Laser-cutting equipment must be totally enclosed to be safe for human operators.* **H:** *Even if the laser machine is fully contained within, there still exist some amount of risk for the workers in the close proximity.*

- ACTIVE/PASSIVE: Whether there is an active-to-passive (or vice versa) transformation from the premise to the hypothesis.
  Example: **P:** *Hani Hanjour, Khalid Al Mihdhar, and Majed Moqed were flagged by capps.* **H:** *Capps never flagged anyone.*

	Annotation Tag	Label Frequency	BiLSTM	Yang	Balazs (S)	Chen (S)
Matched	CONDITIONAL	5%	100%	100%	100%	100%
	WORD_OVERLAP	6%	50%	63%	63%	63%
	NEGATION	26%	71%	75%	75%	75%
	ANTO	3%	67%	50%	50%	50%
	LONG_SENTENCE	20%	50%	**75%**	**75%**	67%
	TENSE_DIFFERENCE	10%	64%	68%	68%	**86%**
	ACTIVE/PASSIVE	3%	75%	75%	75%	88%
	PARAPHRASE	5%	78%	83%	83%	78%
	QUANTITY/TIME_REASONING	3%	50%	50%	50%	33%
	COREF	6%	83%	83%	83%	83%
	QUANTIFIER	25%	64%	59%	59%	**74%**
	MODAL	29%	66%	65%	65%	**75%**
	BELIEF	13%	74%	71%	71%	73%
Mismatched	CONDITIONAL	5%	100%	80%	80%	100%
	WORD_OVERLAP	7%	58%	62%	62%	76%
	NEGATION	21%	69%	73%	73%	72%
	ANTO	4%	58%	58%	58%	58%
	LONG_SENTENCE	20%	55%	67%	67%	67%
	TENSE_DIFFERENCE	4%	71%	71%	71%	89%
	ACTIVE/PASSIVE	2%	82%	82%	82%	91%
	PARAPHRASE	7%	81%	89%	89%	89%
	QUANTITY/TIME_REASONING	8%	46%	54%	54%	46%
	COREF	6%	80%	70%	70%	80%
	QUANTIFIER	28%	70%	68%	68%	**77%**
	MODAL	25%	67%	67%	67%	**76%**
	BELIEF	12%	73%	71%	71%	74%

Table 4: This table shows the accuracy of different models for each tagged subset of our 1,000-example development set sample. The '(S)' indicates that results for the single model are shown. Some results that stand out to us are shown in bold.

- PARAPHRASE: Whether the two sentences are close paraphrases.
  Example: **P:** *Uh, lets see.* **H:** *Let us look.*

- COREF: Whether the hypothesis contains a pronoun or referring expression that needs to be resolved using the premise.
  Example: **P:** *You and I, gentle reader, are accredited members of the guild.* **H:** *We are recognised as members of the guild.*

- QUANTIFIER: Whether either sentence contains one of the following quantifiers: *much, enough, more, most, less, least, no, none, some, any, many, few, several, almost, nearly*.
  Example: **P:** *We have provided an invoice to facilitate your gift.* **H:** *There's no invoice available for your gift.*

- MODAL: Whether either sentence contains one of the following modal verbs: *can, could, may, might, must, will, would, should.*
  Example: **P:** *Conversely, an increase in government saving adds to the supply of resources available for investment and may put downward pressure on interest rates.* **H:** *The amount of resources available for investment increases when government savings are increased.*

- BELIEF: Whether either sentence contains one of the following belief verbs: *know, believe, understand, doubt, think, suppose, recognize, recognize, forget, remember, imagine, mean, agree, disagree, deny, promise*.
  Example: **P:** *I trust that this is a fillip of propaganda and not a serious query.* **H:** *I believe this is to get attention and not a real inquiry.*

- NEGATION: Whether either sentence contains negation.
  Example: **P:** *On reflection, the parts will hold together.* **H:** *The parts will not hold together.*

- ANTO: Whether the two sentences contain an antonym pair.
  Example: **P:** *As united 93 left Newark, the flight's crew members were unaware of the hijacking of American 11.* **H:** *As the flight United 93 left Newark the crew members were fully aware of the hijacking of American 11.*

- TENSE_DIFFERENCE: Whether the two sentences use different tenses on any verbs.
  Example: **P:** *Does she like what she does?* **H:** *Does she like what she is doing?*

- QUANTITY/TIME_REASONING: Whether understanding the pair requires quantity or time reasoning.
  Example: **P:** *The vice chairman joined the conference shortly before 10:00; the secretary, shortly before 10:30.* **H:** *The secretary joined before the vice chairman.*

- WORD_OVERLAP: Whether the two sentences share more than 70% of their tokens.
  Example: **P:** *Let's look for paua shells!* **H:** *Let's look for sticks.*

- LONG_SENTENCE: Whether the premise or hypothesis is longer than 30 or 16 words respectively.
  Example: **P:** *As invested with its dignity, since the seventeenth century just as the crown has been used for the monarch, or the oval office has come to stand for the president of the United States.* **H:** *Nobody in Britain associates the crown with the monarchy.*

Table 4 shows model results on tagged examples for the BiLSTM baseline and for the three systems for which we were able to acquire example-by-example development set results (submission of these results was optional). Among those tags that are frequent enough to yield clearly interpretable numbers, none indicates a subset of the corpus that is dramatically harder or easier for the submitted models than is the corpus overall. This suggests that—as is typical with neural network models—these models do not rely strongly on any particular structural properties of the input texts to the exclusion of others.

We note that the submitted systems that use intra-attention (the three shown) do relatively well on the LONG_SENTENCE and NEGATION tags. This technique likely helps the encoders to recover the structures of long sentences and to correctly identify the scope of instances of negation. We also note that all systems do relatively poorly on the QUANTITY/TIME_REASONING section, suggesting that these simple sentence feature extractors are not well situated to learn quantitative reasoning in this setting.

Authors	1-NN Genre Accuracy
Chen et al.	67.3%
Nie and Bansal	74.0%
Balazs et al.	69.2%
Vu et al.	67.0%
Yang et al.	54.7%

Table 5: A thousand sentences are randomly sampled from the *matched* test set and their pairwise distances to all sentences in the test set (premises and hypotheses) are calculated. This table shows the percentage of times the first nearest neighbor belongs to the same genre as the sample sentence.

6.2 Nearest Neighbors

Test Set Sentences The competition participants were asked to submit sentence vectors for all the premise and hypothesis sentences in the test sets. We randomly sample 1,000 sentences from the *matched* test set and, using cosine similarity, calculate their pairwise distances against all sentences in the *matched* test set. Table 5 shows the percentage of times the first nearest neighbor belongs to the same genre as the chosen sentence. All models score fairly highly on this metric, suggesting that the learned representations are not genre-agnostic, despite their effectiveness in unseen genres. The models with higher percentage accuracy on the NLI task (see Table 3) show better performance on this metric as well, suggesting that this genre clustering property correlates with the overall quality of the metric space that each model uses to represent sentences.

The better models are also more interpretable. Table 6 shows example sentences and their three nearest neighbors for all models. It appears that entity identity is important for the Nie and Bansal model, though not it a way that is tied to syntactic position. For the *Critics loved Merchant-Ivory* example, we see matches to critics. In the *Students love the rich culture* example, we similarly see many matches to school and love. Since for each premise sentence in the MultiNLI corpus, there are 3 associated hypothesis sentences, it's not surprising to see that the first nearest neighbor is often one of these associated sentences, like in the *Critics* example where the first nearest neighbor for all systems is the premise sentence. We found that for some examples, the better performing systems like Nie and Bansal's had all three as-

Sample	Model	Nearest Neighbours	
Students love the rich culture and history of the school. (TR.)	Chen	TEL.	my son loved learning about computers in high school
		TR.	Families love this city-within-a-city on the beach.
		TR.	The urban working class loved the new factories.
	Nie	TEL.	my son loved learning about computers in high school
		TEL.	I really loved it when I was in middle school.
		SL.	A librarian and fellow patient kindled his love for literature more than school.
	Balazs	SL.	School, more than anything else, was credited for his love of literature.
		SL.	A librarian and fellow patient kindled his love for literature more than school.
		TEL.	I really loved it when I was in middle school.
	Vu	SL.	A librarian and fellow patient kindled his love for literature more than school.
		TR.	France's oldest city is a wonderful destination, with rich history and extreme beauty.
		FIC.	The rave had some of the best artists and celebrities.
	Yang	TR.	The urban working class loved the new factories.
		FIC.	my son loved learning about computers in high school
		TR.	This area is a favorite of hikers who enjoy invigorating journeys through dense forests and along the river valleys celebrated in the paintings of Gustave Courbet.
Critics loved Merchant-Ivory's final movie, which was an adaption of a novel written by Kaylie Jones. (SL.)	Chen	SL.	Critics laud Merchant-Ivory's exit from the 19th century in this adaptation of a semiautobiographical novel by Kaylie Jones (daughter of novelist James Jones).
		SL.	I loved Begnigni's movie!
		SL.	Mercer was the lifelong love of Franklin Roosevelt, and the revelation of their affair nearly ended his marriage to Eleanor.
	Nie	SL.	Critics laud Merchant-Ivory's exit from the 19th century in this adaptation of a semiautobiographical novel by Kaylie Jones (daughter of novelist James Jones).
		SL.	Critics find the book entertaining, praising digressions on gambling, laughing, and love, as well as Pinker's pop-culture references.
		SL.	Critics think that Lichtenstein was a contemporary genius.
	Balazs	SL.	Critics laud Merchant-Ivory's exit from the 19th century in this adaptation of a semiautobiographical novel by Kaylie Jones (daughter of novelist James Jones).
		TEL.	The period of the civil war is very interesting to me, I've read about 3 novels about that, including John Jakes ones.
		SL.	The most vivid moments in Kubrick's films in the last 30 years have come when he has turned his actor's faces into Think of Malcolm McDowell in A Clockwork Orange (1971), Jack Nicholson in The Shining (1980), and Vincent D'Onofrio in Full Metal Jacket (1987).
	Vu	SL.	Critics laud Merchant-Ivory's exit from the 19th century in this adaptation of a semiautobiographical novel by Kaylie Jones (daughter of novelist James Jones).
		SL.	Mercer was the lifelong love of Franklin Roosevelt, and the revelation of their affair nearly ended his marriage to Eleanor.
		SL.	Critics find the book entertaining, praising digressions on gambling, laughing, and love, as well as Pinker's pop-culture references.
	Yang	SL.	Critics laud Merchant-Ivory's exit from the 19th century in this adaptation of a semiautobiographical novel by Kaylie Jones (daughter of novelist James Jones).
		TR.	Visitors are encouraged to come during daylight hours, when the park is safer and better patrolled by employees.
		TR.	All Ireland loves a horse, and County Kildare can claim to be at the heart of horse country.

Table 6: Showing the three nearest neighbors for example sentences from a random 1,000-sample subset of the *matched* test set. All results are for single (non-ensemble) models. The genres have been abbreviated.

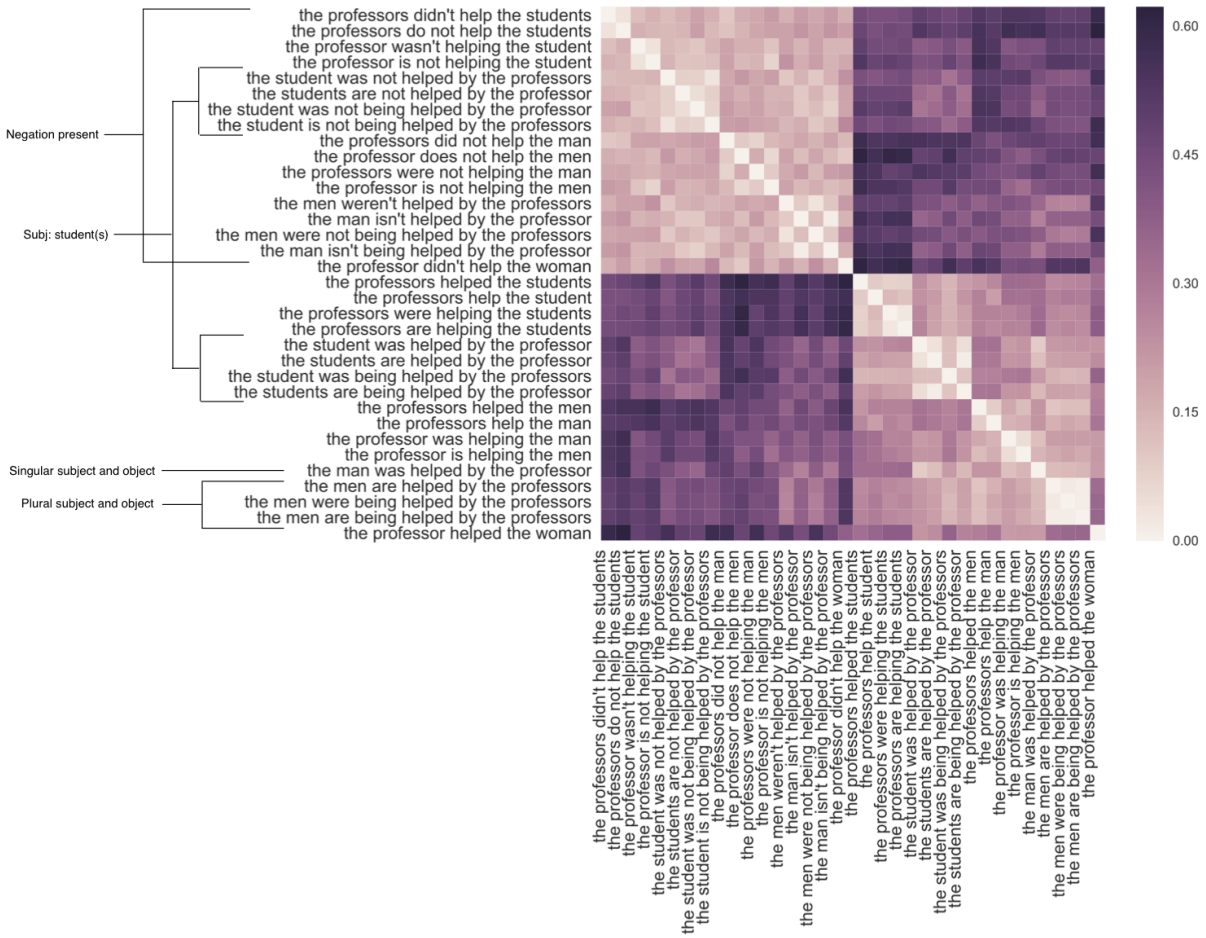

Figure 1: A heatmap showing the cosine similarity between sentence vectors. The vectors were rendered by the Nie and Bansal model. The plots for other systems are very similar.

sociated sentences as their top three nearest neighbors.

Probe Sentences During the competition, we additionally provided a set of automatically generated probe sentences meant to aid error analysis. These probe sentences are produced to vary along dimensions relevant to probing for semantic role and negation information. We asked submitting teams to supply vectors for these sentences in addition to those in the test set. Figure 1 shows the cosine similarity between a subset of these sentence vectors rendered by Nie and Bansal's (2017) system. We find that all systems (except that of Balazs et al., who did not submit these vectors) show similar behavior on these sentences, and we do not observe a clear correlation between behavior here and model performance. Perhaps unsurprisingly, we observe that sentences tend to be more similar to one another the more structural features they have in common. We observe this clearly for negation, identity of the subject, and tense, though continuous tenses are not reliably differentiated from others.

7 Conclusion

We find that BiLSTM-based models with max pooling or intra-sentence attention represent a popular and effective strategy for sentence encoding, and that systems based on this technique perform very well at the task of NLI.

We note that all submitted systems performed reasonably well across the many subsets of the data reflected by our supplementary tags, suggesting that none of these models exploit any particular narrow feature of the task or data to perform well. We also note that model performance does not vary much between the *matched* and *mismatched* sections of the test set. This means that submitted systems are likely capturing reasonably general strategies for extracting representations of meaning from text. As the systems get better, and

fit the training data more closely, the disparity between *matched* and *mismatched* sets may appear. Both of these findings, though, bolster our expectation that the best of the submitted systems represent some of the best general-purpose architectures for sentence encoding available.

However, the task of NLI is far from being solved, and no submitted system approaches human performance, suggesting that there is ample room for further research on both the task and on the more general problem of sentence representation learning. Since many of the examples in MultiNLI require substantial commonsense background knowledge to solve fully, we suspect that the use of large outside datasets and resources (labeled or otherwise) will be crucial to making substantial further progress in this setting.

Acknowledgments

This work was made possible by a Google Faculty Research Award to Sam Bowman and Angeliki Lazaridou, and was also supported by a gift from Tencent Holdings. Allyson Ettinger contributed the supplementary probe sentences. We also thank George Dahl and the organizers of the RepEval 2016 and RepEval 2017 workshops for their help and advice.

References

Jorge Balazs, Edison Marrese-Taylor, Pablo Loyola, and Yutaka Matsuo. 2017. Refining raw sentence representations for textual entailment recognition via attention. In *The Second Workshop on Evaluating Vector Space Representations for NLP (RepEval 2017)*.

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

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

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017b. Recurrent neural network-based sentence encoder with gated attention for natural language inference. In *Proceedings of The Second Workshop on Evaluating Vector Space Representations for NLP (RepEval 2017)*.

Gennaro Chierchia and Sally McConnell-Ginet. 1991. *Meaning and Grammar*. MIT Press, Cambridge, MA.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proc. EMNLP*.

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.

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

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-aware neural language models. In *Proc. AAAI*.

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

Yixin Nie and Mohit Bansal. 2017. Shortcut-stacked sentence encoders for multi-domain inference. In *Proceedings of The Second Workshop on Evaluating Vector Space Representations for NLP (RepEval 2017)*.

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

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. In *Proc. ICLR*.

Hoa Trong Vu, Thuong-Hai Pham, Xiaoyu Bai Marc Tanti, Lonneke van der Plas, and Albert Gatt. 2017. LCT-MALTA's submission to RepEval 2017 shared task. In *Proceedings of The Second Workshop on Evaluating Vector Space Representations for NLP (RepEval 2017)*.

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

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

Han Yang, Marta R. Costa-jussà, and José A. R. Fonollosa. 2017. Character-level intra attention network for natural language inference. In *Proceedings of The Second Workshop on Evaluating Vector Space Representations for NLP (RepEval 2017)*.