Enhancing Natural Language Inference Using New and Expanded Training Data Sets and New Learning Models

Arindam Mitra[*] **Ishan Shrivastava**[*] **Chitta Baral**
Arizona State University
{amitra7,ishrivas,chitta}@asu.edu

Abstract

Natural Language Inference (NLI) plays an important role in many natural language processing tasks such as question answering. However, existing NLI modules that are trained on existing NLI datasets have several drawbacks. For example, they do not capture the notion of entity and role well and often end up making mistakes such as "Peter signed a deal" can be inferred from "John signed a deal". As part of this work, we have developed two datasets that help mitigate such issues and make the systems better at understanding the notion of "entities" and "roles". After training the existing models on the new dataset we observe that the existing models do not perform well on one of the new benchmark. We then propose a modification to the "word-to-word" attention function which has been uniformly reused across several popular NLI architectures. The resulting models perform as well as their unmodified counterparts on the existing benchmarks and perform significantly well on the new benchmarks that emphasize "roles" and "entities".

Introduction

Natural language inference (NLI) is the task of determining the truth value of a natural language text, called "hypothesis" given another piece of text called "premise". The list of possible truth values include *entailment*, *contradiction* and *neutral*. *Entailment* means that the hypothesis must be true if the premise is true. *Contradiction* indicates that the hypothesis can never be true if the premise is true. *Neutral* pertains to the scenario where the hypothesis can be both true and false as the premise does not provide enough information. Table 1 shows an example of each of the three cases.

NLI has many applications in natural language processing and natural language understanding. In particular, it can be used in various natural language question answering domains. For example, in (Mitra et al. 2019) NLI is used in question answering on some of the ARISTO science question answering domains. In one of the example domains of that paper, text (Frog-LC) about life cycle of a Frog is given. With respect to that text questions and possible answer choices are given. One example of that is:

premise: A soccer game with multiple males playing. **hypothesis:** Some men are playing a sport. **label:** entailment.
premise: A man in a black shirt is playing golf outside. **hypothesis:** The man in the black shirt trades Pokemon cards with his girlfriend. **label:** contradiction.
premise: A girl swings high in the air. **hypothesis:** A girl is gaining momentum to flip off the swing **label:** neutral.

Table 1: Example premise-hypothesis pairs from SNLI dataset with human-annotated labels.

Q: What is the middle stage in a frog's life?
(A) tadpole with legs (B) froglet

In that paper NLI is used to answer the above question in the following way. Using the two answer choices two natural language textual hypothesis: "Tadpole with legs is the middle stage in a frog's life" and "Froglet is the middle stage in a frog's life" are constructed. Then NLI is used to compute the degree of entailment between the textual premise Frog-LC and each of the textual hypothesis and answer between (A) and (B) based on the degree of entailments with respect to the corresponding textual hypothesis. Such an approach of using NLI for question answering is very attractive as the alternative of extracting relevant facts from the textual premise Frog-LC, translating the question and answer choices to formal representations, and connecting them has multiple avenues where error can be introduced. Having said that, to make the approach of using NLI for QA more useful one needs better NLI systems. Moreover since most NLI systems are developed using supervised learning over NLI datasets, there is a need for good and comprehensive NLI datasets.

Recently several large scale datasets have been produced to advance the state-of-the-art in NLI. One such dataset is SNLI which contains a total of 570k premise-hypothesis pairs. However, several top performing systems on SNLI struggle when they are subjected to examples which require understanding the notion of entity and semantic roles. Table 2 shows some examples of this kind.

[*] These authors contributed equally to this work.

premise:	John went to the kitchen.
hypothesis:	Peter went to the kitchen.
premise:	Kendall lent Peyton a bicycle.
hypothesis:	Peyton lent Kendall a bicycle.

Table 2: Sample premise-hypothesis pairs where existing models trained on SNLI suffers significantly.

The top-performing models on the SNLI benchmark wrongly predict *entailment* as the correct label for both the examples in Table 2 with very high confidence. For example, the ESIM (Chen et al. 2016) model predicts entailment with a confidence of 82.21% and 96.29% respectively. We observe similar behavior for models that are trained on another well known NLI dataset called the MNLI (Williams, Nangia, and Bowman 2017) dataset.

This is a big concern. For NLI to be better useful NLI modules need to understand roles and entities better. With the use of vector representations distinguishing entities is a challenge, which is somewhat exacerbated by the limitations of the current datasets. Considering the importance of NLI, especially in the context of QA, we address this issue in two ways: (i) by addressing the limitations of the current dataset and (ii) by enhancing existing NLI models so that they can distinguish different entities that may have close vector representations. Our contributions, in this paper, addressing these issues are are twofold: 1) we show how existing annotated corpus such as VerbNet (Schuler 2005), PropBank (Palmer, Gildea, and Kingsbury 2005), QA-SRL (FitzGerald et al. 2018), bAbI (Weston et al. 2015), AMR (Banarescu et al. 2013), CoNLL 2003 Shared(NER) task (Ratinov and Roth 2009), CoNLL 2004 Shared(SRL) task (Carreras and Màrquez 2005) and GMB(Groningen Meaning Bank) (Bos et al. 2017) can be used to automatically create premise-hypothesis pairs that stress on the understanding of entities and roles. 2) We propose a novel neural attention for NLI which combines vector similarity with symbolic similarity to perform significantly better NLI, especially on the new datasets. The code and dataset for this work is available at http://bit.ly/2OwzGgo.

Dataset Generation

In this work we create two new datasets. The first contains examples of *premise-hypothesis* pairs that are labelled as *contradiction* where the *hypothesis* is created from the *premise* by replacing its named entities with a different and disjoint set of named entities. This dataset is referred to as NER-CHANGED. The second one contains examples of *contradiction* labelled *premise-hypothesis* pairs where the hypothesis is created by swapping the two different entities from the *premise* which has the same (VerbNet) type but plays different roles. This one is referred to as the ROLE-SWAPPED. To help the NLI systems learn the importance of these modifications, the two datasets also contain *entailment* labelled *premise-hypothesis* pairs where the *hypothesis* is exactly same as the *premise*.

The two datasets do not contain any *neutral* labelled *premise-hypothesis* pairs. This is because we follow the assumptions made in the creation of the SNLI (Bowman et al. 2015) dataset. The assumption is that the events and entities mentioned in the premise and hypothesis sentences are coreferent. Following this assumption the examples shown in Table 2 are labelled as *contradiction*. We follow this assumption to label the *premise-hypothesis* for the two new datasets.

NER-CHANGED DataSet

To create this data set, we utilize the sentences from the bAbI (Weston et al. 2015) corpus, the AMR (Banarescu et al. 2013), the CoNLL 2003 Shared NER task (Ratinov and Roth 2009) corpus and the GMB(Groningen Meaning Bank) (Bos et al. 2017).

Creation of dataset using bAbI We extract all the 30814 sentences which contains a single person name and the 4770 sentences which contain names of two persons. For all the single name sentences, we replace the name in the sentence with the token **personX** to create a set of template sentences. For example, the sentence *"Mary moved to the hallway."* becomes *"PersonX moved to the hallway."*

This way, we create a total of 398 unique template sentences, each consisting only one name. We then use a list of 15 gender-neutral names to replace the token **PersonX** in all the template sentences. We then make pairs of premise and hypothesis sentences and label the ones with different names as *contradiction* and with same name as *entailment*. The template mentioned above, creates the following *premise-hypothesis* pair:

Premise : Kendall moved to the hallway.
Hypothesis : Peyton moved to the hallway.
Gold Label: contradiction

Similarly, we use the sentences with names of two persons and the gender-neutral names to create more *contradiction* labelled *premise-hypothesis* pairs. We ensure that the set of unique template sentences and gender-neutral names are disjoint for train, dev, and test set.

Creation of dataset using AMR and CoNLL 2003 Shared task (NER) datasets Contrary to the bAbI dataset, the AMR corpus and the CoNLL 2003 Shared task (NER) datasets contain complex and lengthier sentences which provides variety to our dataset. We use the annotations available in the two datasets to extract a total 4855 template sentences such that each of them contains at least one mention of a person or a location (a city or a country). Consider the following examples with the mention of a city (CoNLL and AMR) and a person (AMR):

CoNLL: *"Teheran defied international pressure by announcing plans to produce more fuel for its nuclear program."*
AMR: *"New Delhi subsequently said it regretted the incident, which it said had been the result of a misunderstanding."*
AMR: *"William Perry said the equipment can detect particles down to the picogram (a millionth of a millionth of a gram)."*

We create three lists of certain names of cities, countries and persons respectively selected from the AMR corpus. We use these lists to change the names mentioned in the candidate sentences and create the “contradiction” labelled *premise-hypothesis* pair. From the examples mentioned above, the following pairs are generated:

Premise : Dublin defied international pressure by announcing plans to produce more fuel for its nuclear program.
Hypothesis : Shanghai defied international pressure by announcing plans to produce more fuel for its nuclear program.
Gold Label: contradiction

Premise : Dublin subsequently said it regretted the incident, which it said had been the result of a misunderstanding.
Hypothesis : Shanghai subsequently said it regretted the incident, which it said had been the result of a misunderstanding.
Gold Label: contradiction

Creation of dataset using GMB We also use this corpus to collect sentences containing “Numbers” and “Dates” to create *contradiction* labelled *premise-hypothesis* pairs. This corpus provides both Part-Of Speech(POS) and Named-Entity(NER) annotations that enable us to extract sentences with mentions of different kinds of “Numbers” and “Dates” entity. We identify and extract sentences containing two types of “Numbers” entity 1) “Cardinal in Numerics” 2) “Cardinal in Words”.

The sentences with the mention of at least one token with the NER annotation as “O” and the POS annotation as “CD”(*cardinal number*) is chosen to create the template sentences for “Numbers” entity type. A simple check of whether this token is a number or not further segregates the template sentences into the template sentences for the two types of “Numbers” entity considered in this work. Consider the following example with a mention of a “Numbers” entity of the “Cardinal in Numerics” type:

“Australia has about 5000 troops in Iraq as part of the U.S. led coalition.”

We use two disjoint sets, one for premise and the other for hypothesis sentences. Each consists of thirty random numbers ranging from 10 to 20000 that serves as the replacement options for the “Numbers” entity of “Cardinal in Numerics” type. These sets are used to fill in the template sentences to create the “contradiction” labelled *premise-hypothesis* pairs. For the example mentioned above the following pair is generated:

Premise : “Australia has about 14061 troops in Iraq as part of the U.S. led coalition.’
Hypothesis : “Australia has about 8958 troops in Iraq as part of the U.S. led coalition.”
Gold Label: *contradiction*

Similarly we use two more disjoint sets that contain only 2 digit numbers for premise and hypothesis sentences. These are automatically converted to words to generate the “contradiction” labelled *premise-hypothesis* pairs from the template sentences for “Cardinal in Words” type.

We also identify and extract sentences containing three types of “Dates” entity: 1) “Year” 2) “Month” 3) “Day of the week”.

We shortlist sentences with at least one token with NER annotation as “B-tim”/“I-tim”. If the POS annotation for this token is “CD”(*cardinal number*) and the token is of length four the sentence is considered as the template sentence for “Year” type of “Dates” entity. If the POS annotation for this token is anything else then it can be either a month or a day in the week. A simple token match for the 12 month names and the 7 day names generate the template sentences for “Month” type and “Day of the week” type of “Dates” entity.

We use two disjoint sets, one for premise and the other for hypothesis sentences. Each set consists of twenty, 4 digit numbers ranging from 1900 to 2019 that serve as the replacement options for the template sentences of “Year” type of “Dates” entity. The names of the 12 months and 7 days of the week are used to fill in the the respective type of “Dates” entity template sentences to create the “contradiction” labelled *premise-hypothesis* pairs. An example of such a pair is shown below:

Premise : “The spokesman says a formal agreement on the project will be signed in February when Indonesian President Susilo Bambang Yudhoyono is scheduled to visit Moscow.”
Hypothesis : “The spokesman says a formal agreement on the project will be signed in November when Indonesian President Susilo Bambang Yudhoyono is scheduled to visit Moscow.”
Gold Label: *contradiction*

ROLES-SWITCHED DataSet

The ROLES-SWITCHED dataset contains sentences such as “John rented a bicycle to David”, where two person play two different roles even though they participate in the same event (verb). We use the VerbNet (Schuler 2005) lexicon to extract the set of all verbs (events) that take as arguments two same kinds of entities for two different roles. We use this set to extract annotated sentences from VerbNet (Schuler 2005), PropBank (Palmer, Gildea, and Kingsbury 2005), QA-SRL(FitzGerald et al. 2018) CoNLL 2004 Shared SRL Task (Carreras and Màrquez 2005) and CoNLL 2003 Shared NER Task (Ratinov and Roth 2009), which are then used to create sample *premise-hypothesis* pairs. The following two subsections describe the process in detail.

Creation of dataset using VerbNet VerbNet(Schuler 2005) provides a list of VerbNet class of verbs and also provides the restrictions defining the types of thematic roles that are allowed as arguments. It also provides a list of member verbs for each class of verbs. For example, consider the VerbNet class for the verb give - “give-13.1”. The roles it can take are “Agent”, “Theme” and “Recipient”. It further provides the restrictions as “Agent” and “Recipient” can only be either an Animate or an Organization type of entity.

We use this information provided by VerbNet(Schuler 2005) to shortlist 35 VerbNet classes (verbs) that accepts the same kind of entities for different roles. “give-13.1” is one such class as the two different roles for it, “Agent” and “Recipient” accepts the same kind of entities, namely “Animate” or “Organization”. We take the member verbs from each of the shortlisted VerbNet classes to compute the set of all 646 “interesting” verbs. We then extract the annotated sentences from VerbNet to finally create the template sentences for the data set creation.

Consider the following sentence from VerbNet which contains the verb “lent” which is a member verb of the VerbNet class “give-13.1”.
“They lent me a bicycle.”

We use such sentences and associated annotations to create template sentences such as:
“PersonX lent PersonY a bicycle.”

Note that VerbNet provides example sentence for each VerbNet classes not for individual member verbs and sometimes the example sentence might not contain the required **PersonX** and **PersonY** slot. Thus, using this technique, we obtain a total of 87 unique template sentences from VerbNet. These sentences are very simple and thus can be easily converted to different tenses. We therefore convert all the template sentences into present tense in 3rd person and future tense to expand our list of template sentences. We also use the member verbs as synonyms to finally create 1611 unique templates. For all such template sentences, we use gender-neutral names to create the contradiction labelled role-swapped *premise-hypothesis* pairs, as shown below:

Premise : Kendall lent Peyton a bicycle.
Hypothesis : Peyton lent Kendall a bicycle.
Gold Label: contradiction

Creation of dataset using PropBank PropBank (Proposition Bank) is a large corpus with annotations for propositions and predicate argument relations. It also provides a mapping to VerbNet. We use these mappings to VerbNet in order to extract sentences from PropBank for the shortlisted VerbNet Classes. Not all the extracted sentences are ideal to create the desired template sentences. For example:
“The Beatles give way to baseball in the Nipponese version.”

Therefore we manually remove such sentences to create 13 unique template sentences. An example of one such template sentence is shown below:
“PersonX also is being advised by PersonY.”

Since the template sentences here are more complicated as compared to VerbNet template sentences, we manually convert them into different tenses to create more template sentences. We also use the VerbNet member verbs to expand the list to get 89 unique template sentences. We use the list of gender-neutral names to create the contradiction labelled role-swapped *premise-hypothesis* pairs, as shown below:

Premise : Kendall also is being advised by Peyton.
Hypothesis : Peyton also is being advised by Kendall.
Gold Label: contradiction

Creation of dataset using QA-SRL In the QA-SRL (FitzGerald et al. 2018) dataset, roles are represented as questions. Thus we go through the list of questions from the QA-SRL dataset to map the questions into their corresponding VerbNet role. We consider only those QA-SRL sentences which contains both the role-defining questions of a verb in their annotation and where each of the entity associated with those two roles (the answer to the questions) is either a singular or a plural noun, or a singular or a plural proper noun. We then swap those two entities to create a *contradiction* labelled *premise-hypothesis* pair.

For example, consider the VerbNet class “defend-85” which is shortlisted based on the criteria mentioned in the section 2.1. This class has the verb “protect” as one of its member verbs. We look for all the examples from the QA-SRL dataset that contain the role-defining questions for the verb “protect”. Once such example is shown below:

Sentence : *In Germany, the Emperor had repeatedly* ***protected*** *Henry the Lion against complaints by rival princes or cities especially in the cases of Munich and Lbeck.*
Base Verb : *protect*
Who did someone protect? : *Henry the Lion*
Who protected someone?: *the Emperor*

Based on the exampe mentioned above we swap the answers of the two questions and create the following *contradiction* labelled *premise-hypothesis* pair:

Premise : In Germany, the Emperor had repeatedly protected Henry the Lion against complaints by rival princes or cities especially in the cases of Munich and Lbeck.
Hypothesis : In Germany, Henry the Lion had repeatedly protected the Emperor against complaints by rival princes or cities especially in the cases of Munich and Lbeck.
Gold Label: contradiction

We also manually shortlist 109 QA-SRL(FitzGerald et al. 2018) sentences to automatically create 109 unique template sentences. For example:
“PersonX asked that she be allowed to inform PersonY before the news was released.”

Similar to templates from other corpora, we use gender-neutral names to fill in these templates and create the contradiction labelled role-swapped *premise-hypothesis* pairs, as shown below:

Premise : Kendall asked that she be allowed to inform Peyton before the news was released.
Hypothesis : Peyton asked that she be allowed to inform Kendall before the news was released.
Gold Label: contradiction

Details of creation of Dataset using CoNLL 2004 Shared(SRL) task and CoNLL 2004 Shared(NER) task corpora Both CoNLL 2004 Shared(SRL) task and CoNLL 2004 Shared(NER) task provides sentences with NER annotations. We use these annotations to shortlist roughly 350 sentences with mentions of names of two person entities. We manually filter out sentences that will lead to grammatically incorrect or incoherent sentences after switching the roles. We also use the VerbNet member verbs as synonyms to create a total of 305 unique template sentences. These template sentences are more complex as compared to the VerbNet template sentences. Here’s an example of a template sentence from the CoNLL 2004 ad CoNLL

2003 datasets respectively:

"Col. North conveyed the request to his superiors and to Assistant Secretary of State PersonX, who will deliver it to Secretary of State PersonY."

"PersonX has decided not to endorse PersonY as the presidential candidate of the Reform Party, CNN reported late Tuesday."

Similar to the previous sections we use the gender-neutral names to replace the two entities which are then swapped to create a *contradiction* labelled *premise-hypothesis* pair. The following pairs are created for the examples mentioned above:

Premise : Col. North conveyed the request to his superiors and to Assistant Secretary of State Kendall, who will deliver it to Secretary of State Peyton.
Hypothesis : Col. North conveyed the request to his superiors and to Assistant Secretary of State Peyton, who will deliver it to Secretary of State Kendall.
Gold Label: contradiction

Premise : Kendall has decided not to endorse Peyton as the presidential candidate of the Reform Party, CNN reported late Tuesday..
Hypothesis : Peyton has decided not to endorse Kendall as the presidential candidate of the Reform Party, CNN reported late Tuesday..
Gold Label: contradiction

Model

In this section we describe the proposed modification to the existing attention mechanism of the DecAtt (Parikh et al. 2016) and the ESIM (Chen et al. 2016) model that helps in performing better on the NER CHANGED dataset.

Let *a* be the premise and *b* be the hypothesis with length l_a and l_b such that a = (a_1,a_2,...,a_{l_a}) and b = (b_1,b_2,...,b_{l_b}) where each a_i and $b_j \in \mathrm{R}^{\mathrm{d}}$ is a word vector embedding of dimensions *d*.

Both DecAtt and the ESIM models first transform the original sequence a and b to another sequence $\bar{a} = (\bar{a}_1, ..., \bar{a}_{l_a})$ and $\bar{b} = (\bar{b}_1, ..., \bar{b}_{l_b})$ of same length to learn task-specific word embeddings. They then compute a non normalized attention between each pair of words using dot product as shown in equation 1.

$$e_{\mathrm{ij}} = (\bar{a}_{\mathrm{i}})^T \bar{b}_{\mathrm{j}} \tag{1}$$

Since the initial word embeddings for similar named entities such as "john" and "peter" are very similar, the normalized attention scores between NER-CHANGED sentence pairs such as " Kendall moved to the hallway." and "Peyton moved to the hallway." forms a diagonal matrix which normally occurs when premise is exactly same as hypothesis (Figure 1). As a result, the systems end up predicting *entailment* for this kind of premise-hypothesis pairs. To deal with this issue, we introduce symbolic similarity into the attention mechanism. The attentions scores are then computed as follows:

$$e'_{\mathrm{ij}} = \lambda_{\mathrm{ij}} e_{\mathrm{ij}} + (1 - \lambda_{\mathrm{ij}}) sym_{\mathrm{ij}} \tag{2}$$

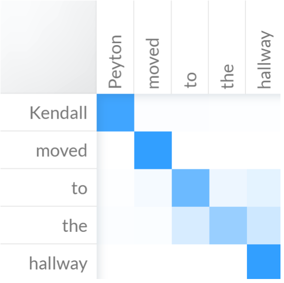

Figure 1: Word-to-word attention computed by the DecAtt model that is trained on the SNLI model.

Here, *sym*$_{ij}$ represents the symbolic similarity which is assigned 0 if the string representing $\mathrm{a_i}$ is not "equal" to the string representing $\mathrm{b_j}$. If the two strings match, then a weight *w* which is a hyper-parameter, is assigned. $\lambda_{ij} \in [0, 1]$ is a learnable parameter which decides how much weight should be given to vector similarity and how much weight to the symbolic similarity (*sym*$_{ij}$) while calculating the new unnormalized attention weights *e'*$_{ij}$. λ_{ij} is computed using equation 3. We will refer to this feed-forward neural network as the lambda layer.

$$\lambda_{ij} = 1 - LReLU(1 - LReLU(W_\lambda x_{ij}^\lambda)) \tag{3}$$

Equation 3 ensures that $\lambda_{\mathrm{ij}} \in [0, 1]$. The LReLU refers to the Leaky Rectified Linear Unit (Maas, Hannun, and Ng 2013). Here, W_λ is learned from data with respect to the NLI task and x_{ij}^λ is the input to the lambda layer which is a 16 dimensional sparse feature vector and encodes the NER (Named Entity Recognition) information for the pair of words in the two sentences. We group the NER information into 4 categories namely 'Name", "Numeric", "Date" and "Other". We use Spacy and Stanford NER tagger to obtain the NER category of a word. Let v_i^{ner} and v_j^{ner} be two vectors in $\{0, 1\}^4$ which encode the one-hot representation of the NER category, then $x_{ij}^\lambda[k_1 * 4 + k_2] = v_i^{ner}[k_1] * v_j^{ner}[k_2]$ where k_1 and $k_2 \in \{0, 1, 2, 3\}$.

Related Work

Many large labelled NLI datasets have been released so far. Bowman et al. develop the first large labelled NLI dataset containing $570k$ premise-hypothesis pairs. They show sample image captions to crowd-workers and the label (entailment, contradiction and neutral) and ask workers to write down a hypothesis for each of those three scenarios. As a result they obtain a high agreement entailment dataset known as Stanford Natural Language Inference (SNLI). Since premises in SNLI contains only image captions it might contain sentences of limited genres. MNLI (Williams, Nangia, and Bowman 2017) have been developed to address this issue. Unlike SNLI and MultiNLI, SciTail (Khot, Sabharwal, and Clark 2018) and QNLI (Demszky, Guu, and Liang 2018) consider multiple-choice question-answering as an NLI task to create the SciTail and QNLI datasets respectively. Recent datasets like PAWS (Zhang, Baldridge, and He 2019) which is a paraphrase identification dataset

Exp Id	**Data Sets**		**DecAtt**		**ESIM**		**Lambda DecAtt**		**Lambda ESIM**		**BERT**	
	Train	Test	Train Acc(%)	Test Acc(%)	Train Acc(%)	Test Acc(%)	Train Acc(%)	Test Acc(%)	Train Acc(%)	Test Acc(%)	Train Acc(%)	Test Acc(%)
1	SNLI	NC	84.58	33.80	89.78	**56.11**	85.10	53.69	90.10	53.95	88.45	55.56
2	SNLI + NC	NC	88.52	82.91	91.21	69.77	88.56	97.81	92.14	**99.13**	85.19	69.31
3	SNLI + NC	SNLI	88.52	83.50	91.21	85.09	88.56	82.94	92.14	87.10	85.19	**88.26**
4	SNLI	RS	84.58	49.96	89.78	50.46	85.10	49.98	90.11	**54.14**	88.45	49.74
5	SNLI + RS	RS	78.96	49.93	89.66	**93.72**	78.98	49.94	90.23	90.11	82.33	49.78
6	SNLI + RS	SNLI	78.96	85.31	89.66	86.99	78.98	84.52	90.23	87.70	82.33	**88.47**
7	SNLI + RS + NC	NC	84.18	80.72	92.63	75.64	84.77	95.32	92.75	**98.91**	81.26	69.00
8	SNLI + RS + NC	SNLI	84.18	83.71	92.63	87.03	84.77	84.24	92.75	87.28	81.26	**88.08**
9	SNLI + RS + NC	RS	84.18	50.11	92.63	84.45	84.77	50.08	92.75	**87.92**	81.26	49.80
10	MNLI	NC	74.47	61.16	83.71	70.35	74.00	**78.76**	85.20	68.15	81.87	54.19
11	MNLI + NC	NC	84.40	85.56	88.82	75.58	83.50	**97.94**	88.30	95.12	80.13	68.03
12	MNLI + NC	MNLI MisM	84.40	71.49	88.82	75.59	83.50	70.08	88.30	74.29	80.13	79.96
13	MNLI + NC	MNLI M	84.40	71.76	88.82	76.75	83.50	69.95	88.30	75.17	80.13	79.74
14	MNLI	RS	74.47	50.08	83.71	50.16	74.00	50.12	85.20	50.64	81.87	50.18
15	MNLI + RS	RS	69.75	50.08	85.01	50.51	63.40	50.13	84.12	50.12	75.09	49.53
16	MNLI + RS	MNLI MisM	69.75	71.58	85.01	75.75	63.40	70.85	84.12	74.51	75.09	80.90
17	MNLI + RS	MNLI M	69.75	71.72	85.01	76.65	63.40	71.03	84.12	74.65	75.09	80.56
18	MNLI + RS + NC	NC	74.90	60.25	90.09	75.33	78.30	96.17	89.79	91.91	76.91	68.53
19	MNLI + RS + NC	RS	74.90	50.08	90.09	51.18	78.30	69.87	89.79	53.35	76.91	50.27
20	MNLI + RS + NC	MNLI MisM	74.90	64.37	90.09	75.45	78.30	69.97	89.79	75.72	76.91	80.75
21	MNLI + RS + NC	MNLI M	74.90	64.56	90.09	77.29	78.30	50.11	89.79	76.48	76.91	80.74

Table 3: Table shows the train and test set accuracy for all the experiments. Here, NC refers to NER-Changed dataset, RS refers to the Role-Switched dataset, MNLI MisM refers to MNLI Mismatched test set and MNLI M refers to MNLI Matched test set. Each row of this table represents an experiment. The Second and Third columns of each row represents the train set and the test set used for that experiment. Rest of the columns show the train and the test accuracy (Acc) in percentages for all the five models. In our experiments, we have used the *bert-large-uncased* model.

and EQUATE (Ravichander et al. 2019) which evaluates quantitative reasoning in natural language inference also helps to advance the field of NLI. Glockner, Shwartz, and Goldberg creates a NLI test set which shows the inability of the current state of the art systems to accurately perform inference requiring lexical and world knowledge.

Since the release of such large data sets, many advanced deep learning architectures have been developed (Bowman et al. 2016; Vendrov et al. 2015; Mou et al. 2015; Liu et al. 2016; Rocktäschel et al. 2015; Wang and Jiang 2015; Cheng, Dong, and Lapata 2016; Parikh et al. 2016; Munkhdalai and Yu 2016; Paria et al. 2016; Chen et al. 2016; Khot, Sabharwal, and Clark 2018; Devlin et al. 2018; Liu et al. 2019). Although many of these deep learning models achieve close to human level performance on SNLI and MultiNLI datasets, these models can be easily deceived by simple adversarial examples. Kang et al. shows how simple linguistic variations such as negation or re-ordering of words deceives the DecAtt Model. Gururangan et al. goes on to show that this failure is attributed to the bias created as a result of crowd sourcing. They observe that crowd sourcing generates hypothesis that contain certain patterns that could help a classifier learn without the need to observe the premise at all.

Experiments and Analysis

We split the NER-Changed and Role-Switched dataset in train/dev/test sets each containing respectively 289K/26.5k/26.6K and 129K/8.5k/9k premise-hypothesis pairs, which is then used to evaluate the performance of a total of five models. This includes three existing models, namely DecAtt (Parikh et al. 2016), ESIM (Chen et al. 2016) and BERT (Devlin et al. 2018) and our two new models namely Lambda DecAtt and Lambda ESIM. We use the 300 dimensional GloVe(Pennington, Socher, and Manning 2014) embeddings to represent the input tokens for DecAtt, ESIM, Lambda DecAtt and Lambda ESIM models in all of our experiments. The results are shown in Table 3.

Row $1, 4, 10$ and 14 shows that if the models are trained on the SNLI train set or MNLI train set alone, they perform poorly on the NER-CHANGED and ROLE-SWITCHED test set . For e.g., row 1 shows that the highest performance that is achieved on the NE-changed test data after training on the SNLI train data is 56.11% (ESIM model). This shows that the knowledge of roles and entities that are provided through the developed datasets is missing in the SNLI and the MNLI datasets.

We also experiment by combining the SNLI train set individually with the two new datasets and train the 5 models. Rows $2, 3, 5\&6$ shows those results. As shown in row 2 after exposing the NER-CHANGED train set at train time along with the SNLI training dataset, DecAtt shows some improvement where the ESIM and BERT models continue to struggle in the NER-CHANGED test set. On the other hand, as shown in row 5, when we expose the ROLE-SWITCHED train set at train time along with the SNLI training dataset, ESIM shows significant improvement where the DecAtt and BERT models continue to struggle in the ROLE-SWITCHED test set. Our Lambda DecAtt and Lambda ESIM models however significantly outperform the remaining models on the NER-CHANGED test set and achieves as well as or better accuracy than its unmodified counterparts DecAtt and ESIM on the SNLI test set.

We also train the 5 models by combining the train sets of SNLI and the two new datsets. Rows $7, 8, \&9$ shows those results. When we expose both the NER-CHANGED and ROLE-SWITCHED train sets at train time along with the SNLI train set, our Lambda ESIM model comes out to be the best performing model as compared to rest of the 4 models on the NER-CHANGED and ROLE-SWITCHED test sets. It also achieves a better accuracy than its unmodified counterpart on the SNLI test set. Our Lamda DecAtt model gives comparable performance to our Lambda ESIM model on the NER-CHANGED test set but continues to suffer on the ROLE-SWITCHED test set. This behavior is also seen between the original DecAtt and ESIM models. Both the original ESIM and our Lambda ESIM model, perform a BiLSTM based transformation over the input embedding. The lack of such a transformation in the original DecAtt and our Lambda DecAtt model suggests that this could be the reason behind their poor performance on the ROLE-SWITCHED test set.

For experiments with the MNLI dataset instead of the SNLI dataset, we observe the same behavior on the NER-CHANGED test set. However we observe that the performance on the ROLE-SWITCHED test set is always significantly better when combining the ROLE-SWITCHED train set with the SNLI train set instead of MNLI train set.

Weight Vector Dimensions	Dimension Meaning	Lambda DecAtt (Learnt Weight)	Lambda ESIM (Learnt Weight)
1	“Names-Names”	**0.5418**	**0.342**
2	“Names-Dates”	0.309	0.353
3	“Names-Num”	0.2571	0.164
4	“Names-Others”	0.713	0.374
5	“Date-Names”	0.409	0.511
6	“Dates-Dates”	**0.359**	**0.287**
7	“Dates-Num”	0.374	0.466
8	“Dates-Others”	0.566	0.350
9	“Num-Names”	0.474	0.501
10	“Num-Dates”	0.522	0.413
11	“Num-Num”	**0.635**	**0.444**
12	“Num-Others”	0.522	0.351
13	“Others-Names”	0.528	0.378
14	“Others-Dates”	0.709	0.327
15	“Others-Num”	0.243	0.325
16	“Others-Others”	**0.869**	**0.372**

Figure 2: Learnt weights by the Lambda Layer for Lambda DecAtt and Lambda ESIM models when trained on SNLI.

Weight Vector Dimensions	Dimension Meaning	Lambda DecAtt (Learnt Weight)	Lambda ESIM (Learnt Weight)
1	“Names-Names”	**-0.022**	**0.138**
2	“Names-Dates”	0.312	0.040
3	“Names-Num”	0.483	0.282
4	“Names-Others”	0.616	0.513
5	“Date-Names”	0.439	0.098
6	“Dates-Dates”	**-0.032**	**-0.123**
7	“Dates-Num”	0.470	0.330
8	“Dates-Others”	0.715	0.525
9	“Num-Names”	0.484	0.296
10	“Num-Dates”	0.400	0.144
11	“Num-Num”	**0.310**	**0.394**
12	“Num-Others”	0.558	0.478
13	“Others-Names”	0.607	0.393
14	“Others-Dates”	0.690	0.465
15	“Others-Num”	0.468	0.302
16	“Others-Others”	**0.811**	**0.451**

Figure 3: Learnt weights by the Lambda Layer for Lambda DecAtt and Lambda ESIM models when trained on SNLI and NER Changed. A higher value indicates more weight being given to Vector Similarity, while a smaller value indicates more weight being given to Symbolic Similarity

Figure 2 and 3 compare the lambda values for the scenario when Lambda DecAtt and Lambda ESIM models are trained on only SNLI to the scenario when they are trained on SNLI and NER-CHANGED together. Recall that a higher value indicates more weight being given to vector similarity, while a smaller value indicates more weight being given to symbolic similarity. Figure 2 and 3 shows that with the lambda layers the NLI models are giving more priority to symbolic similarity while matching name-name, number-number or date-date pairs.

Conclusion

We have shown how the existing meaning representation datasets can be used to create NLI datasets which stress on the understanding of entities and roles. Furthermore, we show that popular existing models when trained on existing datasets hardly understand the notion of entities and roles. We have proposed a new attention mechanism for natural language inference. As experiments suggest, the new attention function significantly helps to capture the notion of entities and roles. Furthermore, the performance on the existing testbeds does not drop with the new attention mechanism.

Acknowledgement

Support from DARPA, and NSF grant 1816039 is acknowledged.

References

Banarescu, L.; Bonial, C.; Cai, S.; Georgescu, M.; Griffitt, K.; Hermjakob, U.; Knight, K.; Koehn, P.; Palmer, M.; and Schneider, N. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, 178–186.

Bos, J.; Basile, V.; Evang, K.; Venhuizen, N.; and Bjerva, J. 2017. *The Groningen Meaning Bank*. 463–496.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. *CoRR* abs/1508.05326.

Bowman, S. R.; Gauthier, J.; Rastogi, A.; Gupta, R.; Manning, C. D.; and Potts, C. 2016. A fast unified model for parsing and sentence understanding. *CoRR* abs/1603.06021.

Carreras, X., and Màrquez, L. 2005. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, CONLL '05, 152–164. Stroudsburg, PA, USA: Association for Computational Linguistics.

Chen, Q.; Zhu, X.; Ling, Z.; Wei, S.; and Jiang, H. 2016. Enhancing and combining sequential and tree LSTM for natural language inference. *CoRR* abs/1609.06038.

Cheng, J.; Dong, L.; and Lapata, M. 2016. Long short-term memory-networks for machine reading. *CoRR* abs/1601.06733.

Demszky, D.; Guu, K.; and Liang, P. 2018. Transforming question answering datasets into natural language inference datasets. *CoRR* abs/1809.02922.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR* abs/1810.04805.

FitzGerald, N.; Michael, J.; He, L.; and Zettlemoyer, L. 2018. Large-scale QA-SRL parsing. *CoRR* abs/1805.05377.

Glockner, M.; Shwartz, V.; and Goldberg, Y. 2018. Breaking NLI systems with sentences that require simple lexical inferences. *CoRR* abs/1805.02266.

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S. R.; and Smith, N. A. 2018. Annotation artifacts in natural language inference data. *CoRR* abs/1803.02324.

Kang, D.; Khot, T.; Sabharwal, A.; and Hovy, E. H. 2018. Adventure: Adversarial training for textual entailment with knowledge-guided examples. *CoRR* abs/1805.04680.

Khot, T.; Sabharwal, A.; and Clark, P. 2018. SciTail: A textual entailment dataset from science question answering. In *AAAI*.

Liu, Y.; Sun, C.; Lin, L.; and Wang, X. 2016. Learning natural language inference using bidirectional LSTM model and inner-attention. *CoRR* abs/1605.09090.

Liu, X.; He, P.; Chen, W.; and Gao, J. 2019. Multi-task deep neural networks for natural language understanding. *CoRR* abs/1901.11504.

Maas, A. L.; Hannun, A. Y.; and Ng, A. Y. 2013. Rectifier nonlinearities improve neural network acoustic models. In *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*.

Mitra, A.; Clark, P.; Tafjord, O.; and Baral, C. 2019. Declarative question answering over knowledge bases containing natural language text with answer set programming. In *AAAI*.

Mou, L.; Men, R.; Li, G.; Xu, Y.; Zhang, L.; Yan, R.; and Jin, Z. 2015. Recognizing entailment and contradiction by tree-based convolution. *CoRR* abs/1512.08422.

Munkhdalai, T., and Yu, H. 2016. Neural Tree Indexers for Text Understanding. *arXiv e-prints* arXiv:1607.04492.

Palmer, M.; Gildea, D.; and Kingsbury, P. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics* 31(1):71–106.

Paria, B.; Annervaz, K. M.; Dukkipati, A.; Chatterjee, A.; and Podder, S. 2016. A neural architecture mimicking humans end-to-end for natural language inference. *CoRR* abs/1611.04741.

Parikh, A. P.; Täckström, O.; Das, D.; and Uszkoreit, J. 2016. A decomposable attention model for natural language inference. *CoRR* abs/1606.01933.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Ratinov, L., and Roth, D. 2009. Design Challenges and Misconceptions in Named Entity Recognition. In *Proc. of the Conference on Computational Natural Language Learning (CoNLL)*.

Ravichander, A.; Naik, A.; Rose, C.; and Hovy, E. 2019. Equate: A benchmark evaluation framework for quantitative reasoning in natural language inference. *arXiv preprint arXiv:1901.03735*.

Rocktäschel, T.; Grefenstette, E.; Hermann, K. M.; Kočiský, T.; and Blunsom, P. 2015. Reasoning about Entailment with Neural Attention. *arXiv e-prints* arXiv:1509.06664.

Schuler, K. K. 2005. *Verbnet: A Broad-coverage, Comprehensive Verb Lexicon*. Ph.D. Dissertation, Philadelphia, PA, USA. AAI3179808.

Vendrov, I.; Kiros, R.; Fidler, S.; and Urtasun, R. 2015. Order-Embeddings of Images and Language. *arXiv e-prints* arXiv:1511.06361.

Wang, S., and Jiang, J. 2015. Learning natural language inference with LSTM. *CoRR* abs/1512.08849.

Weston, J.; Bordes, A.; Chopra, S.; Rush, A. M.; van Merriënboer, B.; Joulin, A.; and Mikolov, T. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

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

Zhang, Y.; Baldridge, J.; and He, L. 2019. PAWS: Paraphrase Adversaries from Word Scrambling. *arXiv e-prints* arXiv:1904.01130.