Understanding Roles and Entities: Datasets and Models for Natural Language Inference

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

Abstract

We present two new datasets and a novel attention mechanism for Natural Language Inference (NLI). Existing neural NLI models, even though when trained on existing large datasets, 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”. The two datasets have been developed to mitigate such issues and make the systems better at understanding the notion of “entities” and “roles”. After training the existing architectures on the new dataset we observe that the existing architectures does 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 architectures perform as well as their unmodified counterparts on the existing benchmarks and perform significantly well on the new benchmark for “roles” and “entities”.

1 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 the hypothesis must be true as 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.

Recently several large scale datasets have been produced to advance the state-of-the-art in NLI.

∗ These authors contributed equally to this work.

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.

The top-performing models on the SNLI benchmark 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.

To help the NLI systems to better learn these concepts of entity and semantic roles we present two new datasets. Our contributions are twofold: 1) we show how existing annotated corpus such as VerbNet (Schuler, 2005), AMR (Banarescu et al., 2013) and QA-SRL(FitzGerald et al., 2018) 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 on the new datasets.

2 Dataset Generation

We create two new datasets. The first one contains examples of neutral or contradiction labelled premise-hypothesis pairs 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 neutral 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 ROLE-SWAPPED. To help the NLI systems to 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.

2.1 NER-CHANGED DataSet

To create this data set, we utilize the sentences from the bAbI (Weston et al., 2015) and the AMR (Banarescu et al., 2013) corpus.

2.1.1 Creation of examples using bAbI

We extract all the 40814 sentences which contain a single person name and the 4770 sentences which contain two person names. 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 following 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 neutral 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 | Neutral |

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

2.1.2 Creation of examples using AMR

Contrary to the bAbI dataset, AMR corpus contains complex and lengthier sentences which provides variety to our dataset. We use the annotation available in the AMR corpus to extract 945 template sentences such that each of them contain at least one mention of a city or a country or a person. Consider the following example with the mention of a city:

“Teheran defied international pressure by announcing plans to produce more fuel for its nuclear program.”

Using a list of certain names of cities, countries and persons selected from the AMR corpus we change the names mentioned in the candidate sentences to create the “Neutral” labelled premise-hypothesis pair. From the example mentioned above, the following pair is 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 | Neutral |

We also use the AMR corpus to collect sentences containing “Numbers” and “Dates” to create neutral or contradiction labelled premise-hypothesis pair. The gold labels in this case is decided manually. The following pair provides an example of this case:

| Premise | The Tajik State pays 35 dirams (a few cents) per day for every person in the rehabilitation clinics. |
|---------|-------------------------------------------------------------------------------------------------------------------|
| Hypothesis | The Tajik State pays 62 dirams (a few cents) per day for every person in the rehabilitation clinics. |
| Gold Label | Contradiction |

We also convert a few numbers to their word format and replace them in the sentences to create premise-hypothesis pairs. Consider the following example:

| Premise | The Tajik State pays 35 dirams (a few cents) per day for every person in the rehabilitation clinics. |
|---------|-------------------------------------------------------------------------------------------------------------------|
| Hypothesis | The Tajik State pays 62 dirams (a few cents) per day for every person in the rehabilitation clinics. |
| Gold Label | Contradiction |

We also use the two name sentences and the gender neutral names to create more neutral labelled premise-hypothesis pairs. We ensure that the set of unique template sentences and gender neutral names are disjoint for train, dev, test set.
2.2 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) and QA-SRL(FitzGerald et al., 2018), which are then used to create sample premise-hypothesis pairs. The following two subsections describe the process in detail.

2.2.1 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 47 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 89 unique template sentences from VerbNet. For all such template sentences, we use gender neutral names to create the neutral labelled role-swapped premise-hypothesis pairs, as shown below:

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

2.2.2 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(FitzGerald et al., 2018) dataset to map the questions into their corresponding VerbNet role. We consider only those QA-SRL(FitzGerald et al., 2018) 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 neutral labelled premise-hypothesis pair. The following pair shows an example:

Premise: Many kinds of power plant have been used to drive propellers.
Hypothesis: Propellers have been used to drive many kinds of power plant.
Gold Label: neutral

3 Model

In this section we describe the existing attention mechanism of the DecAtt (Parikh et al., 2016) and the ESIM (Chen et al., 2016) model. We then describe the proposed modification which helps to perform 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 $b_j \in \mathbb{R}^d$ is a word vector embedding of dimensions $d$.

Both DecAtt and the ESIM model first transforms the original sequence $a$ and $b$ to another se-
Table 3: Table shows the train and test set accuracy for all the experiments. Here, NC refers to NER-CHANGED dataset and RS refers to the ROLE-SWITCHED dataset. Each row of this table represents an experiment. The first two 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.

\[
\bar{a}_{ij} = (\bar{a}_1, ..., \bar{a}_{n_t}) \text{ and } \bar{b} = (\bar{b}_1, ..., \bar{b}_{n_r}) \text{ of same length to learn task-specific word embeddings.
}
\]

It then computes a non normalized attention between each pair of words using dot product as shown in equation 1.

\[
e_{ij} = (\bar{a}_j)^T \bar{b}_j
\]  
(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. 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:

\[
e'_{ij} = \lambda_{ij}e_{ij} + (1 - \lambda_{ij})sym_{ij}
\]  
(2)

Here, symmetric represents the symbolic similarity which is assigned 0 if the string representing \(a_i\) is not “equal” to the string representing \(b_j\). If the two strings matches, 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 \(e'_{ij}\). \(\lambda_{ij}\) is computed using equation 3. We will refer to this feed-forward neural network as the lambda layer.

\[
\lambda_{ij} = 1 - \text{LeakyReLU}(W_{\lambda}x_{ij})
\]  
(3)

Here, \(W_{\lambda}\) is learned from data with respect to the NLI task and \(x_{ij}\) 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 \(v^{ner}_{i}\) and \(v^{ner}_{j}\) be two vectors in \(\{0, 1\}^3\) which encode the one-hot representation of the NER category, then \(x_{ij}^{\lambda}[k_1 * 4 + k_2] = v^{ner}_{i}[k_1] * v^{ner}_{j}[k_2]\) where \(k_1\) and \(k_2\) ∈ \(\{0, 1, 2, 3\}\).

4 Related Works

Many large labelled NLI datasets have been released so far. Bowman et al. (2015) 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. MultiNLI (Williams et al., 2017) have been developed to address this issue. Unlike SNLI and MultiNLI, (Khot et al., 2018) and (Demszky et al., 2018) considers multiple-choice question answering as an NLI task to create the SciTail (Khot et al., 2018) and QNLI (Demszky et al., 2018) datasets respectively. Recent datasets like PAWS (Zhang et al., 2019) which is a paraphrase identification dataset also helps to advance the field of NLI. Glockner et al. (2018) 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; Munkhdalai and Yu, 2016; Rocktäschel et al., 2015; Wang and Jiang, 2015; Cheng et al., 2016; Parikh et al., 2016; Munkhdalai and Yu, 2016; Sha et al., 2016; Paria et al., 2016; Chen et al., 2016; Khot et al., 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. (2018) shows how simple linguistic variations such as negation or re-ordering of words deceives the DecAtt Model. Gururangan et al. (2018) 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.

5 Experiments

We split the NER-CHANGED and ROLE-SWITCHED dataset in train/dev/test sets each containing respectively 85.7K/4.4k/4.2k and 10.4/1.2k/1.2k 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 two new models namely Lambda DecAtt (ours) and Lambda ESIM (ours). The results are shown in Table 3.

We observe that if the models are trained on the SNLI train set alone, they perform poorly on the NER-CHANGED and ROLE-SWITCHED test set. This could be attributed to the absence of similar examples in the SNLI dataset. After exposing the new datasets at train time along with the SNLI training dataset, DecAtt and BERT shows significant improvement where the ESIM model continues to struggle in the NER-CHANGED test set. Our Lambda DecAtt and Lambda ESIM models however significantly outperform the remaining models and achieves as well as or better accuracy than its unmodified counterparts DecAtt and ESIM on the SNLI test set.

6 Conclusion

We have shown how the existing annotated 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 does not drop on the existing testbeds when the new attention mechanism is used, which shows the generality of the proposed attention mechanism.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pages 178–186.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. CoRR, abs/1508.05326.

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. CoRR, abs/1603.06021.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, and Hui Jiang. 2016. Enhancing and combining sequential and tree LSTM for natural language inference. CoRR, abs/1609.06038.
Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long short-term memory-networks for machine reading. CoRR, abs/1601.06733.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. CoRR, abs/1809.02922.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805.

Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. 2018. Large-scale QA-SRL parsing. CoRR, abs/1805.05377.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. CoRR, abs/1805.02266.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. CoRR, abs/1803.02324.

Dongyep Kang, Tushar Khot, Ashish Sabharwal, and Eduard H. Hovy. 2018. Adventure: Adversarial training for textual entailment with knowledge-guided examples. CoRR, abs/1805.04680.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In AAAI.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. CoRR, abs/1901.11504.

Yang Liu, Chengjie Sun, Lei Lin, and Xiaolong Wang. 2016. Learning natural language inference using bidirectional LSTM model and inner-attention. CoRR, abs/1605.09090.

Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2015. Recognizing entailment and contradiction by tree-based convolution. CoRR, abs/1512.08422.

Tsendsuren Munkhdalai and Hong Yu. 2016. Neural semantic encoders. CoRR, abs/1607.04315.

Tsendsuren Munkhdalai and Hong Yu. 2016. Neural Tree Indexers for Text Understanding. arXiv e-prints, page arXiv:1607.04492.

Biswajit Paria, K. M. Amnervaz, Ambedkar Dukkipati, Ankush Chatterjee, and Sanjay Podder. 2016. A neural architecture mimicking humans end-to-end for natural language inference. CoRR, abs/1611.04741.

Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. CoRR, abs/1606.01933.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiský, and Phil Blunsom. 2015. Reasoning about Entailment with Neural Attention. arXiv e-prints, page arXiv:1509.06664.

Karin Kipper Schuler. 2005. Verbnet: A Broad-coverage, Comprehensive Verb Lexicon. Ph.D. thesis, Philadelphia, PA, USA. AAI3179808.

Lei Sha, Baobao Chang, Zhifang Sui, and Sujuan Li. 2016. Reading and thinking: Re-read lstm unit for textual entailment recognition. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2870–2879, Osaka, Japan. The COLING 2016 Organizing Committee.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2015. Order-Embeddings of Images and Language. arXiv e-prints, page arXiv:1511.06361.

Shuohang Wang and Jing Jiang. 2015. Learning natural language inference with LSTM. CoRR, abs/1512.08849.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. arXiv preprint arXiv:1502.05698.

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

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase Adversaries from Word Scrambling. arXiv e-prints, page arXiv:1904.01130.