Exploring End-to-End Differentiable Natural Logic Modeling

Yufei Feng[†]*, **Zi'ou Zheng**[†]*, **Quan Liu**[‡], **Michael Greenspan**[†], **Xiaodan Zhu**[†]
[†]Electrical and Computer Engineering & Ingenuity Labs Research Institute, Queen's University
[‡]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research
{feng.yufei,ziou.zheng,michael.greenspan,xiaodan.zhu}@queensu.ca
quanliu@iflytek.com

Abstract

We explore end-to-end trained differentiable models that integrate natural logic with neural networks, aiming to keep the backbone of natural language reasoning based on the natural logic formalism while introducing subsymbolic vector representations and neural components. The proposed model adapts module networks to model natural logic operations, which is enhanced with a memory component to model contextual information. Experiments show that the proposed framework can effectively model monotonicity-based reasoning, compared to the baseline neural network models without built-in inductive bias for monotonicity-based reasoning. Our proposed model shows to be robust when transferred from upward to downward inference. We perform further analyses on the performance of the proposed model on aggregation, showing the effectiveness of the proposed subcomponents on helping achieve better intermediate aggregation performance.

1 Introduction

A recent research trend has attempted to further advance the long-standing problem of bringing together the complementary strengths of neural networks and symbolic models, e.g., the research performed in (Garcez et al., 2015; Yang et al., 2017; Rocktäschel and Riedel, 2017; Evans and Grefenstette, 2018; Weber et al., 2019; De Raedt et al., 2019; Mao et al., 2019), among others. It is known that neural models can approximate complex functions and are robust to noise and ambiguity, while symbolic models often render superior explainability and interpretability but are brittle and prone to fail in the presence of noise and uncertainty.

The majority of research efforts are based on some abstract logical forms such as the first-order logic (FOL) or its *fragments*. For natural language, obtaining such a representation is known to face many thorny challenges. Natural logic instead aims to sidestep some of the challenges by performing inferences over surface forms of text based on *monotonicity* or *projectivity* (Van Benthem, 1986; Valencia, 1991; MacCartney and Manning, 2009; Icard and Moss, 2014), and has been applied to tasks such as natural language inference (MacCartney and Manning, 2009; Angeli and Manning, 2014) and question answering (Angeli et al., 2016).

In this work we explore differentiable natural logic models that integrate natural logic with neural networks, with the aim to keep the backbone of inference based on the natural logic formalism, while introducing subsymbolic vector representations and neural components into the framework. Combining the advantages of neural networks with natural logic needs to take several basic problems into consideration. Two problems flow directly from this objective: 1) How (and where) to leverage the strength of neural networks in the natural logic formalism, and; 2) How to alleviate the issue of a lack of intermediate supervision for training sub-components, which may lead to the spurious problem (Guu et al., 2017; Min et al., 2019) in the end-to-end training.

We explore a framework in which module networks (Andreas et al., 2016; Gupta et al., 2020) are leveraged to model the natural logic operations, which is enhanced with a memory module component to

*Equal contribution.

Proceedings of the 28th International Conference on Computational Linguistics, pages 1172–1185
Barcelona, Spain (Online), December 8-13, 2020

Relation	Relation Name	Example	Set Theoretic Definition
$x \equiv y$	equivalence	$mom \equiv mother$	$x = y$
$x \sqsubset y$	forward entailment	$cat \sqsubset animal$	$x \subset y$
$x \sqsupset y$	reverse entailment	$animal \sqsupset cat$	$x \supset y$
$x \,^\wedge\, y$	negation	$human \,^\wedge\, nonhuman$	$x \cap y = \varnothing \wedge x \cup y = U$
$x \mid y$	alternation	$cat \mid dog$	$x \cap y = \varnothing \wedge x \cup y \neq U$
$x \smile y$	cover	$animal \smile nonhuman$	$x \cap y \neq \varnothing \wedge x \cup y = U$
$x \# y$	independence	$happy \# student$	all other cases

Table 1: Seven natural logic relations proposed by MacCartney and Manning (2009).

Quantifier	Projection	Input Relation r						
		$\equiv$	$\sqsubset$	$\sqsupset$	$\wedge$	$\mid$	$\smile$	$\#$
all	$\rho^{arg1}(r)$	$\equiv$	$\sqsupset$	$\sqsubset$	$\mid$	$\#$	$\mid$	$\#$
	$\rho^{arg2}(r)$	$\equiv$	$\sqsubset$	$\sqsupset$	$\mid$	$\mid$	$\#$	$\#$
some	$\rho^{arg1}(r)$	$\equiv$	$\sqsubset$	$\sqsupset$	$\smile$	$\#$	$\smile$	$\#$
	$\rho^{arg2}(r)$	$\equiv$	$\sqsubset$	$\sqsupset$	$\smile$	$\#$	$\smile$	$\#$
no	$\rho^{arg1}(r)$	$\equiv$	$\sqsupset$	$\sqsubset$	$\mid$	$\#$	$\mid$	$\#$
	$\rho^{arg2}(r)$	$\equiv$	$\sqsupset$	$\sqsubset$	$\mid$	$\#$	$\mid$	$\#$

Table 2: The projection function ρ can project an input relation r into a different relation depending on the context. Here we show the projection function for each argument position for quantifier *all*, *some* and *no*.

$\bowtie$	$\equiv$	$\sqsubset$	$\sqsupset$	$\wedge$	$\mid$	$\smile$	$\#$
$\equiv$	$\equiv$	$\sqsubset$	$\sqsupset$	$\wedge$	$\mid$	$\smile$	$\#$
$\sqsubset$	$\sqsubset$	$\sqsubset$	$\#$	$\mid$	$\mid$	$\#$	$\#$
$\sqsupset$	$\sqsupset$	$\#$	$\sqsupset$	$\smile$	$\#$	$\smile$	$\#$
$\wedge$	$\wedge$	$\smile$	$\mid$	$\equiv$	$\sqsupset$	$\sqsubset$	$\#$
$\mid$	$\mid$	$\#$	$\mid$	$\sqsubset$	$\#$	$\sqsubset$	$\#$
$\smile$	$\smile$	$\smile$	$\#$	$\sqsupset$	$\sqsupset$	$\#$	$\#$
$\#$	$\#$	$\#$	$\#$	$\#$	$\#$	$\#$	$\#$

Table 3: Relation aggregation table (Icard, 2012). Relations listed in the first column are aggregated with those listed in the first row, yielding the relations in the corresponding entries in the table.

capture contextual information. At the lexical and local relation learning layers, we constrain the network to predict the seven natural logic relations. The entire model is differentiable and end-to-end trained.

We evaluate and analyze the proposed model on the monotonicity subset of Semantic Fragments (Richardson et al., 2020), HELP (Yanaka et al., 2019b) and MED (Yanaka et al., 2019a). We also extend MED to generate a dataset to help evaluate 2-hop inference. The model can effectively learn natural logic operations in the end-to-end training paradigm.[1]

2 Related Work

2.1 Neural Symbolic Models

A growing number of research efforts have recently revisited the long-standing problem of bringing together the complementary advantages of neural networks and symbolic methods. There are at least two approaches that have received intensive attention. One uses symbolic constraints as regularizers to equip neural models with the corresponding inductive bias (Demeester et al., 2016; Diligenti et al., 2017; Donadello et al., 2017; Xu et al., 2018; Li and Srikumar, 2019). Another approach develops differentiable end-to-end trained frameworks based on symbolic models. For example, the work in (Rocktäschel and Riedel, 2017; Weber et al., 2019; Minervini et al., 2020) proposes a differentiable backward-chaining algorithm, and Dong et al. (2019) adopt probabilistic tensor representations for logic predicates and mimic the forward-chaining proof. Evans and Grefenstette (2018) treat inductive logic programming as a satisfiability problem and Manhaeve et al. (2018) combine high-level symbolic oriented reasoning with low-level neural perception models. The second approach is more interesting to us for exploring powerful reasoning models with built-in explainability. Unlike the existing work based on abstract logical forms, this paper explores the integration of neural networks with natural logic.

[1]Our code is available at `https://github.com/feng-yufei/Neural-Natural-Logic`

2.2 Natural Logic

Natural logic (Lakoff, 1970; van Benthem, 1988; Valencia, 1991; Van Benthem, 1995; Nairn et al., 2006; MacCartney, 2009; Icard, 2012; Angeli et al., 2016) has a long history that is traceable to the syllogisms of Aristotle. It aims to model a subset of logical inferences by operating directly on the surface form and structure of language, based on monotonicity or projectivity (Van Benthem, 1986; Valencia, 1991; MacCartney and Manning, 2009; Icard and Moss, 2014), rather than deduction on the abstract forms such as the first-order logic (FOL) or its fragments—it is well known that deriving logic forms for natural language is a very challenging task.

In natural language processing, the framework proposed in (MacCartney and Manning, 2008; MacCartney and Manning, 2009) extends monotonicity-based models (van Benthem, 1988; Valencia, 1991) to incorporate semantic exclusion and unifies them to consider implicatives (Nairn et al., 2006), which is a state-of-the-art natural logic formalism that has been used for multiple NLP tasks (MacCartney, 2009; Angeli and Manning, 2014). In this work we explore neural natural logic based on this formalism. We will briefly review the background in Section 3.

2.3 Natural Language Inference

Previous work often studies natural logic in natural language inference (NLI). NLI (Dagan et al., 2005; Iftene and Balahur-Dobrescu, 2007; MacCartney and Manning, 2008; MacCartney and Manning, 2009; MacCartney, 2009; Angeli and Manning, 2014; Bowman et al., 2015), also known as recognizing textual entailment (RTE), aims to model the logical relationships between two sentences, e.g., as a binary (*entailment* vs. *non-entailment*) or three-way classification (*entailment*, *contradiction*, and *neutral*). Recently deep learning algorithms have been proposed (Bowman et al., 2015; Chen et al., 2017a; Chen et al., 2017b; Chen et al., 2017c; Chen et al., 2018; Peters et al., 2018; Yoon et al., 2018; Kiela et al., 2018; Talman et al., 2018; Yang et al., 2019; Devlin et al., 2019). In this paper we will describe and evaluate our neural natural logic models on NLI. The proposed model may also be further extended to other tasks in which natural logic has been applied, e.g., question answering (Angeli et al., 2016).

3 Background

This section briefly reviews the natural logic formalism (MacCartney and Manning, 2009) that our work is based on. For more details, we refer readers to (MacCartney and Manning, 2008; MacCartney and Manning, 2009; MacCartney, 2009; Angeli et al., 2016).

Monotonicity is a pervasive feature of natural language and an essential concept in natural logic (Van Benthem, 1986; Valencia, 1991; MacCartney and Manning, 2009; Icard and Moss, 2014). Similar to the monotone functions in calculus, in natural language upward monotone keeps the entailment relation when the argument "increases" (e.g., *some cats are playing* $\sqsubset$ *some animals are playing*, where *cats* is replaced by its hypernym *animals*). Downward monotone keeps the entailment relation when the argument "decreases" (e.g., *all animals are playing* $\sqsubset$ *all cats are playing*, where *animals* is replaced by its hyponym *cats*).

To extend the monotonicity to consider exclusion, MacCartney and Manning (2009) investigate all sixteen equivalence classes of *set relations* and remove nine degenerate, semantically vacuous relations, thereby defining a seven-relation set $\mathfrak{B} = \{ \equiv, \sqsubset, \sqsupset, \wedge, \mid, \smile, \# \}$ for natural logic, as shown in Table 1.

From a high-level perspective, the natural logic proof system proposed by MacCartney and Manning (2009) consists of the following steps. First, the alignment between two text spans (often two sentences) is obtained and then lexical relation recognition is performed for aligned pairs of words. Consider a simplified example: a premise *All animals outside are eating* and a corresponding hypothesis *All cats outside are playing*, as shown in Figure 1. Each pair of aligned words is assigned one of the relations in Table 1, e.g., *animals* $\sqsupset$ *cats* and *eating* $\mid$ *playing*.

Projection $\rho \colon \mathfrak{B} \to \mathfrak{B}$ is then performed according to the projectivity in specific context. The projection operation has been implemented in the Stanford *natlog* parser[2]. For a given sentence, *natlog* can output the projections at each word position. For example, Table 2 summarizes the projections in the context of

[2]https://stanfordnlp.github.io/CoreNLP/natlog.html

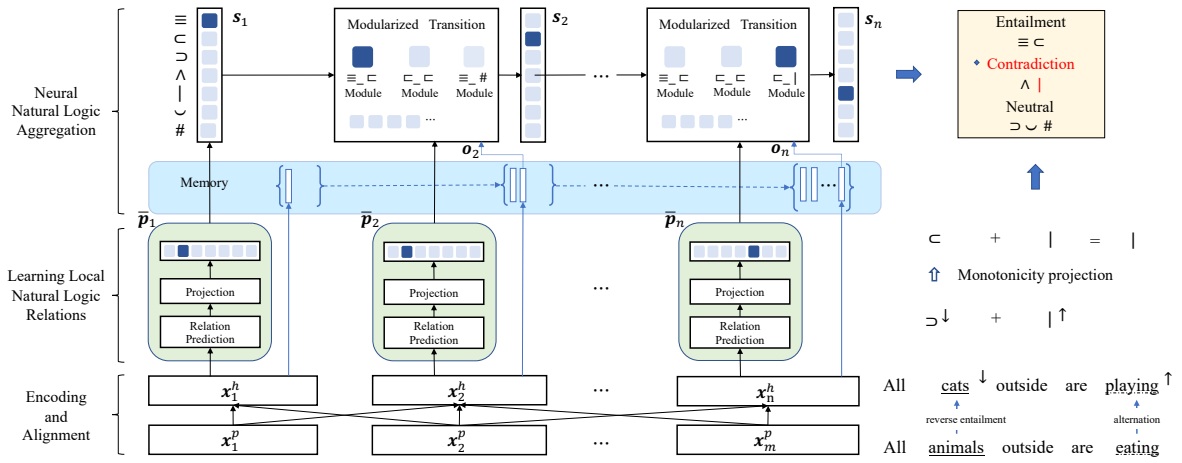

Figure 1: A high-level view of the proposed neural natural logic model.

the quantifier *all*, *some*, and *no*. Specifically, consider the example we discussed in the last paragraph: as *animals* and *cats* take place in the first argument of the quantifier *all*, according to the projectivity in Table 2, the *reverse entailment* relation (*animals* $\sqsupset$ *cats*) will be projected to *forward entailment* (*animals* $\sqsubset$ *cats*) in this specific context. As another example, since *eating* and *playing* take place in the second argument of *all*, the *alternation* relation (*eating* | *playing*) is projected to *alternation* (*eating* | *playing*).

Built on this, relation aggregation is performed to aggregate multiple projected local relations, according to Table 3, to determine the global relation between the sentence pair. In our example, two projected relations, *forward entailment* ($\sqsubset$) and *alternation* (|), are aggregated to yield *alternation* (|); i.e., we obtain *All animals outside are eating* | *All cats outside are playing*. The seven natural logic relationships at the sentence level can be used to determine NLI relations. For example, if NLI is defined as a three-way classification problem (*entailment*, *contradiction*, and *neutral*). The '$\equiv$' or '$\sqsubset$' relation will be mapped to *entailment*, the '$\wedge$' or '|' relation will be mapped to *contradiction*, and '$\sqsupset$', '$\smile$', or '#' to *neutral*.

4 Neural Natural Logic Model

We present a differentiable framework in which natural logic is integrated with neural networks. The overall architecture of the model is shown in Figure 1. At the core of the framework are natural logic operations modeled with memory-enhanced module networks, which are trained end-to-end to optimize the following objective:

$$p(y|\boldsymbol{X}) = \sum_{\boldsymbol{z}\in\mathcal{Z}} p(y|\boldsymbol{z})p(\boldsymbol{z}|\boldsymbol{X}) \tag{1}$$

where y is the output, which in natural language inference is the label of the relation between a premise and hypothesis sentence (e.g., *entailment*, *contradiction*, and *neutral*), and which can be different labels in other tasks. The input $\boldsymbol{X} = \langle \boldsymbol{X}^p, \boldsymbol{X}^h \rangle$ comprises a premise sentence $\boldsymbol{X}^p$ and a hypothesis sentence $\boldsymbol{X}^h$. We use $\boldsymbol{z} = \{z_1, z_2, ..., z_n\}$ to denote a sequence of latent variables corresponding to the output of natural logic aggregation at each time step, where n is the number of hidden variables. The term $\mathcal{Z}$ denotes the space of all possible trajectories and $\boldsymbol{z} \in \mathcal{Z}$. Specifically, for the example in Figure 1, if we perform the aggregation from left to right, $z_1 =$ '$\equiv$', $z_2 = z_3 = z_4 =$ '$\sqsubset$', and $z_5 =$ '|' is a $\boldsymbol{z}$ trajectory that proves the *contradiction* label. Note that $z_i \in \mathfrak{B}$ where $\mathfrak{B}$ is the set of seven relations listed in Table 1.

4.1 Encoding and Alignment

Recent research has shown the effectiveness of distributed representations for encoding lexicons and their semantic relations. We use word embedding and neural networks to learn lexical representations to capture natural logic related semantics. Let $\boldsymbol{X}^p = \{\boldsymbol{x}_1^p, \boldsymbol{x}_2^p, ..., \boldsymbol{x}_m^p\}$ be a premise sentence and

$\boldsymbol{X}^h = \{ \boldsymbol{x}_1^h, \boldsymbol{x}_2^h, ..., \boldsymbol{x}_n^h \}$ the corresponding hypothesis sentence, where m and n are the number of word tokens in the premise and hypothesis, repspectively. Each sentence is fed into a multi-layer BiLSTM, for which $\boldsymbol{a}_i = \mathrm{BiLSTM}(\boldsymbol{X}^p, i)$ denotes the i^{th} hidden vector at the top layer of the BiLSTM, encoding the i^{th} token and its context in the premise. Similarly, we use $\boldsymbol{b}_j = \mathrm{BiLSTM}(\boldsymbol{X}^h, j)$ to denote the hidden vector at the j^{th} position at the top layer of the BiLSTM that encodes the hypothesis. In this paper, we focus on understanding neural natural logic itself, without being further confounded by different ways of exploring knowledge external to the training data, e.g., via pretraining.

Many models can be used to capture cross-sentence attention. Focusing on the training data, the approach proposed in (Chen et al., 2017b) has been widely used in the NLI literature as a baseline. We follow the work to compute cross-sentence attention weight $e_{ij} = \boldsymbol{a}_i^T \boldsymbol{b}_j$ for each pair $\langle \boldsymbol{a}_i, \boldsymbol{b}_j \rangle$. Specifically, for each $\boldsymbol{b}_j$ in the hypothesis, the corresponding content in a premise is weighted summed as $\tilde{\boldsymbol{b}}_j = \sum_{i=1}^{m} \frac{\exp(e_{ij})}{\sum_{k=1}^{m} \exp(e_{kj})} \boldsymbol{a}_i$, which will be used together with $\boldsymbol{b}_j$ to learn local lexical-level inference relations (refer to (Chen et al., 2017b) for more details).

In addition, we compute a *hard alignment indicator* ϕ_j, and $\phi_j = 1$ *if and only if* $\boldsymbol{x}_{i^*}^p = \boldsymbol{x}_j^h$, where $i^* = \arg\max_{i \in \{1,...,m\}} e'_{ij}$.[3] That is, for each word token $\boldsymbol{x}_j^h$ in the hypothesis, we record the token $\boldsymbol{x}_{i^*}^p$ in the premise that has the maximum attention value e'_{ij}. If the word token $\boldsymbol{x}_{i^*}^p$ and $\boldsymbol{x}_j^h$ are the same word type, we let $\phi_j = 1$, which will be used to help reduce the search space in aggregation.

4.2 Learning Local Natural Logic Relation

Given a sequence of alignment $\{ \langle \tilde{\boldsymbol{b}}_1, \boldsymbol{b}_1 \rangle, ..., \langle \tilde{\boldsymbol{b}}_j, \boldsymbol{b}_j \rangle, ..., \langle \tilde{\boldsymbol{b}}_n, \boldsymbol{b}_n \rangle \}$, we use a bi-linear model to compute each pair's probabilistic distribution $\boldsymbol{p}_j$ over the natural logic relations $\mathfrak{B}$:

$$\boldsymbol{p}_j = \mathrm{softmax}(f_s(\tilde{\boldsymbol{b}}_j, \boldsymbol{b}_j)) = \mathrm{softmax}(\tilde{\boldsymbol{b}}_j^T \mathcal{M}^T \boldsymbol{b}_j) \tag{2}$$

In the scoring function f_s, each type of relation $k \in \mathfrak{B}$ has its own weight matrix $\mathcal{M}_k \in \mathbb{R}^{d \times d}$, which is a slice of the tensor $\mathcal{M} \in \mathbb{R}^{d \times d \times |\mathfrak{B}|}$, where d is the dimensionality of $\boldsymbol{b}_j$ or $\tilde{\boldsymbol{b}}_j$. We use *softmax* to normalize the values to be a distribution over $\mathfrak{B}$. Among several alternatives we used, the bi-linear model achieves the best performance on the development dataset, and we use it in our final framework.

4.2.1 Local Relation Constraints

Same as in many other weakly supervised setups, we do not have direct supervision signals here to learn logic relationships at the lexical level; instead, the supervision signals are backpropagated from the overall sentence-level NLI errors. To reduce the search space and alleviate the *spurious* problem (Guu et al., 2017) in which incorrect local inference relationships and aggregation produce correct sentence-level NLI labels,[4] we adopt several strategies as follows.

Symmetric Inference Parameter Sharing: We make the *forward entailment* ($\sqsubset$) and *reverse entailment* ($\sqsupset$) relations share the same parameters. Specifically, to compute $p_j^{\sqsupset}$, we reverse the order of $\langle \tilde{\boldsymbol{b}}_j, \boldsymbol{b}_j \rangle$ to reuse $\mathcal{M}_{\sqsubset}^T$ in the following scoring function, where $\mathcal{M}_{\sqsubset}^T$ is a matrix in $\mathcal{M}^T$ that corresponds to the *forward entailment* ($\sqsubset$) relation.

$$f_s^{\sqsupset}(\tilde{\boldsymbol{b}}_j, \boldsymbol{b}_j) = f_s^{\sqsubset}(\boldsymbol{b}_j, \tilde{\boldsymbol{b}}_j) = \boldsymbol{b}_j^T \mathcal{M}_{\sqsubset}^T \tilde{\boldsymbol{b}}_j \tag{3}$$

Equivalence Constraint: A token pair will be assigned the *equivalence* relation ($\equiv$), if ϕ_j learned above in the alignment stage takes the value of 1:

$$\text{if } \phi_j = 1, \text{ we let } p_j^{\equiv} = 1 \tag{4}$$

[3] Here e'_{ij} is the cross-attention weight obtained from the ESIM model (Chen et al., 2017b) trained on SNLI.

[4] In an extreme case, if a model predicts the first aligned word pair between a premise and hypothesis to be a relation that is consistent with the ground-truth NLI label at the sentence level, the model can choose to ignore all other pairs that follow, and make the correct sentence-level prediction by using the first pair prediction only, even if the aggregation sequence $\boldsymbol{z}$ is incorrect.

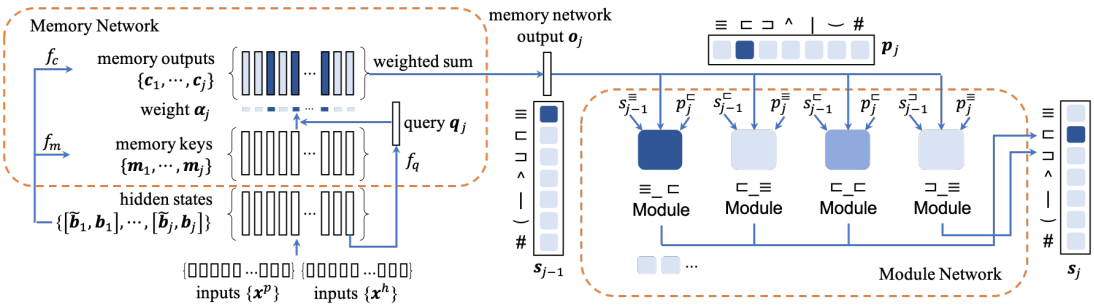

Figure 2: A memory-enhanced module network for natural logic aggregation.

Collapse Constraints: We suppress the relations *negation* ($\wedge$) and *cover* ($\smile$):

$$p_j^{\wedge} = 0, \quad p_j^{\smile} = 0 \tag{5}$$

Inspired by Angeli and Manning (2014), we suppress the *negation* relation ($\wedge$) because its behavior is almost same as that of *alternation* ($|$) in natural logic aggregation, as shown in Table 3, avoiding the co-linearity problem when training on datasets without double negation samples. We also suppress the *cover* relation ($\smile$) because it is extremely rare in current natural language inference datasets.

4.2.2 Projected Distribution

With the predicted seven-dimensional probability vector $\boldsymbol{p}_j$ being ready, our model uses a projection operator ρ to re-organize the distribution according to the projectivity of the corresponding input hypothesis word at position j. Unlike the discrete "hard" projection used in the conventional natural logic, e.g., projecting the first argument of *all* from *reverse entailment* to *forward entailment*, we apply "soft" projection over relation probability distribution $\bar{\boldsymbol{p}}_j$. Specifically, based on the projection Table 2, we convert the original probability distribution $\boldsymbol{p}_j$ to the projected distribution $\bar{\boldsymbol{p}}_j$:

$$\bar{p}_j^{k'} = \sum_k p_j^k \mathbb{1}(\rho(k) = k'), \tag{6}$$

where $\mathbb{1}(\cdot)$ is the indicator function, k is the original relation, and k' is the projected relation. Consider the pair of sentences in Figure 1 and suppose the pair *eating* vs. *playing* have a probability of 0.8 to be *alternation* ($|$) and 0.1 to be *negation* ($\wedge$). According to the projectivity of the second argument of the quantifier *all* in Table 2, both relations are projected to alternation ($|$): $\rho^{playing}(|) = \rho^{playing}(\wedge) = |$. So after projection, $\bar{p}_5^{|} = p_5^{|} + p_5^{\wedge} = 0.9$, where the subscript 5 is the index of the word token *playing* in the hypothesis.

4.3 Aggregation

We propose to leverage the module networks (Andreas et al., 2016; Gupta et al., 2020) to perform neural natural logic aggregation, which is enhanced by a memory network component to leverage the powerful ability in modeling context. Figure 2 shows the proposed neural natural logic aggregation network. The right part of the figure is the aggregation module network and the left is the memory network component.

Specifically, at each time step j, our aggregation algorithm computes a distribution $p(z_j|\boldsymbol{X}) = \mathrm{softmax}(\boldsymbol{s}_j)$, where $\boldsymbol{s}_j = \{s_j^k\}$ is a set of logits. s_j^k is the one corresponding to $p(z_j = k|\boldsymbol{X})$ for relation $k \in \mathfrak{B}$. Our model computes s_j^k with Equation 7.

$$
\begin{aligned}
s_j^k &= \sum_{u\in\mathfrak{B}}\sum_{v\in\mathfrak{B}} G^{u\bowtie v}(s_{j-1}^u, \bar{p}_j^v, \boldsymbol{o}_j)\mathbb{1}(u\bowtie v = k) \\
&= \sum_{u\in\mathfrak{B}}\sum_{v\in\mathfrak{B}} [s_{j-1}^u \cdot \bar{p}_j^v \cdot g^{u\bowtie v}(\boldsymbol{o}_j)]\mathbb{1}(u\bowtie v = k) \\
&= \sum_{u\in\mathfrak{B}} s_{j-1}^u \sum_{v\in\mathfrak{B}} \bar{p}_j^v \cdot g^{u\bowtie v}(\boldsymbol{o}_j)\mathbb{1}(u\bowtie v = k), \qquad (7)
\end{aligned}
$$

At time step 1, $\boldsymbol{s}_1$ is initialized with $\bar{\boldsymbol{p}}_1$. At any other time step $t > 1$, we invoke modules $G^{u\bowtie v}(\cdot)$ to derive $\boldsymbol{s}_j$. Specifically, in our network each relation aggregation in Table 3, i.e., $u \bowtie v$ ($u, v \in \mathfrak{B}$), has its own module $G^{u\bowtie v}(\cdot)$. Now, given the previous $\boldsymbol{s}_{j-1}$ and the current *projected local relation* distribution $\bar{\boldsymbol{p}}_j$, $\boldsymbol{s}_j$ can be computed by marginalizing the Cartesian product $\boldsymbol{s}_{j-1}\cdot\bar{\boldsymbol{p}}_j^T$ according to aggregation Table 3. More specifically, we first compute the Cartesian product $\boldsymbol{s}_{j-1}\cdot\bar{\boldsymbol{p}}_j^T$, which is weighted by the memory $g^{u\bowtie v}(\boldsymbol{o}_j)$. Then for all modules with output being the same relation k according to Table 3, the modules' output are summed up, where $\mathbb{1}(\cdot)$ is the indicator function.

Below we discuss how the memory network response $\boldsymbol{o}_j$ is calculated. In this paper, we propose a memory network component (Weston et al., 2014; Sukhbaatar et al., 2015) to enhance our module aggregation network, aiming to better model contextual information. The details are shown in the left part of Figure 2. Specifically, at time step j, we store memory vectors $\{\boldsymbol{m}_1, ..., \boldsymbol{m}_j\}$ and the corresponding output vectors $\{\boldsymbol{c}_1, ..., \boldsymbol{c}_j\}$ in the memory. The query vector $\boldsymbol{q}_j$ scans the memory and computes the match between itself and memory vectors by taking the inner product followed by a *softmax*:

$$
\begin{aligned}
\boldsymbol{q}_j &= f_q([\tilde{\boldsymbol{b}}_j, \boldsymbol{b}_j]) && (8)\\
\boldsymbol{m}_j &= f_m([\tilde{\boldsymbol{b}}_j, \boldsymbol{b}_j]) && (9)\\
\boldsymbol{c}_j &= f_c([\tilde{\boldsymbol{b}}_j, \boldsymbol{b}_j]) && (10)\\
\alpha_{j,t} &= \mathrm{softmax}(\boldsymbol{q}_j^T\boldsymbol{m}_t),\ t = 1, ..., j && (11)
\end{aligned}
$$

The query, memory, and output vectors are functions of aligned token representation $[\tilde{\boldsymbol{b}}_j, \boldsymbol{b}_j]$, typically modeled by two feed-forward layers. The response vector $\boldsymbol{o}_j$ is computed by the weighted sum over stored outputs vectors $\boldsymbol{c}_j$ and is used in the module network discussed above:

$$
\boldsymbol{o_j} = \sum_{t=1}^{j} \alpha_{j,t} c_t \qquad (12)
$$

where $\boldsymbol{o}_j$ encodes all historical transitions and their context and is then incorporated into Equation 7.

In addition to the sequential aggregation we discuss above in which we perform aggregation left-to-right over a premise and hypothesis pair, we also perform the aggregation on the binarized constituency parses, where aggregation is performed on a tree structure. For node j in the constituency tree, we define a random variable z_j which represents the reasoning states upon seeing the node j and sub-tree, and we use $\boldsymbol{s}_j$ to denote the distribution of z_j. We initialize $\boldsymbol{s}_j$ with projected relation distribution $\bar{\boldsymbol{p}}_j$ if node j is the leaf node. Iteratively, the distribution $\boldsymbol{s}_j$ for each non-leaf node is computed by aggregating its left child (lc) and right child (rc):

$$
s_j^k = \sum_{u\in\mathfrak{B}}\sum_{v\in\mathfrak{B}} G^{u\bowtie v}(s_{lc}^u, s_{rc}^v, \boldsymbol{o}_j)\mathbb{1}(u\bowtie v = k) \qquad (13)
$$

where $\boldsymbol{o}_j$ is the memory network response vector which is computed on the information of all nodes that have already been visited.

Objective Function The final prediction of sentence relation is computed with the distribution of hidden state $\boldsymbol{s}_n$ at the last time step (or the root node if reasoning is performed over the constituency tree). We follow the work of Angeli and Manning (2014) and group relation *equivalence* ($\equiv$) and *forward entailment* ($\sqsubset$) to be *entailment*; *negation* ($\wedge$) and *alternation* ($|$) to be *contradiction*, and; *reverse entailment* ($\sqsupset$), *cover* ($\smile$) and *independent* ($\#$) to be *neutral*. We apply a variant of hard-EM training method (Min et al., 2019), which selects the most likely relation: $p_{entailment} = \max(s_n^{\equiv}, s_n^{\sqsubset})$, $p_{contradiction} = \max(s_n^{\wedge}, s_n^{|})$, and $p_{neutral} = \max(s_n^{\sqsupset}, s_n^{\smile}, s_n^{\#})$. After applying softmax, we obtain the prediction probability, which can be used to compute the cross entropy loss.

5 Experiments

5.1 Setup

Data: We use three datasets that are designed for studying monotonicity based reasoning, i.e., HELP (Yanaka et al., 2019b), MED (Yanaka et al., 2019a), and the monotonicity subset of Semantic Fragments (Richardson et al., 2020). The HELP dataset has 35,891 inference pairs, which are automatically generated by conducting lexical substitution or deletion on one sentence to obtain the other, given natural logic polarity information of each word token and syntactic structure of sentences. The MED dataset contains 5,382 human-generated inference pairs by either asking crowdworkers to perform the generation or manually collecting the pairs from linguistics publications. The monotonicity subset of Semantic Fragments is automatically generated with a controlled set of rules and lexicons, which contains around 2,000 pairs. Since the pairs with the contradiction relation in the Semantic Fragments dataset are obtained by changing quantifiers, which are out of the scope of the natural logic formalism that we use, we do not include this subset in our experiments.

In addition, we create a new *2-hop* dataset. The above datasets lack ground-truth labels for evaluating aggregation at each time step, and most of them are *1-hop* aggregation in which a premise and hypothesis differs only by one span of text. In our *2-hop* dataset, the premise and hypothesis differs by two edits of word/phrase insertion, deletion, or substitution. Our dataset provides ground-truth aggregation output $\{ z_1, ..., z_j, ... z_n \}$ to help assess models' performance on natural logic operations and understand their decision paths. The development of this 2-hop dataset includes three steps: (a) identify the editing type for each example in MED and determine the logic relations; (b) add one more *hop* of relation, and; (c) record the ground-truth aggregation labels at each time step and the final NLI labels following MacCartney's natural logic formalism. We manually checked a subset of the data and found more than 96% of examples are correct. Details of the data development are included in Appendix A.

Implementation Details: Following Chen et al. (2017b), hidden vectors in our model are 300 dimensional. We use pretrained 300-dimensional 840B GloVe vectors (Pennington et al., 2014) to initialize our word embeddings. All word embeddings are trainable after being initialized. We apply a dropout rate of $p = 0.5$. Adam (Kingma and Ba, 2015) is used as our optimizer, and the first momentum is set to be 0.9 and the second 0.999. The batch size is set to be 32 and the initial learning rate is 0.0004. We train ESIM and our neural natural logic models for 32 epochs and use the development set to select models for testing. We use default hyper-parameters specified in (Devlin et al., 2019) and train the BERT-base model for 3 epochs.

5.2 Results

Inference Performance: Table 4 shows the test accuracy of different models on the four datasets that are designed specifically for evaluating monotonicity-based inference. Following Richardson et al. (2020) and Yanaka et al. (2019a), we train the models on SNLI (Bowman et al., 2015) and test on these different test sets. The proposed models, in general, achieve better performances on these four datasets than ESIM (Chen et al., 2017b) and BERT (Devlin et al., 2019). The difference is more prominent in the 2-hop dataset, which requires the system to have a better aggregation ability to make the final prediction.

To demonstrate how the models generalize between the upward and downward monotone, we train the models with HELP's upward monotone subset and test on the downward monotone subset. A system that

Model	Monotonicity Fragments (%)	HELP (%)	MED (%)	Natural Logic 2-Hop (%)	HELP Dev (%) Up Mono.	HELP Test (%) Down Mono.
ESIM	66.18	55.27	51.78	45.13	95.25	21.49
BERT-base	50.58	51.40	45.88	49.33	98.63	13.71
Neural Nat. Log. (seq.)	66.03	58.23	**52.47**	**60.14**	91.20	63.08
Neural Nat. Log. (tree)	**66.47**	**63.95**	47.57	59.97	90.62	**70.80**

Table 4: Test accuracy of the models.

	Model	Precision	Recall	F1
(1)	**Neural Nat. Log. (seq.)**	**0.54**	**0.49**	**0.51**
(2)	(1) w/o. **Memory / Module**	0.49	0.46	0.47
(3)	(2) w/o. **Local Rel. Constraints**	0.12	0.15	0.13

Table 5: Evaluation of models' aggregation performance on the 2-hop dataset.

can better model monotonicity should achieve more robust performance. Specifically, we split the upward monotone subset of the HELP dataset into the training set ($\sim$ 6k training examples) and the development set ($\sim$ 1.5k examples). We train all models on the training split and select models with the highest development accuracy. We test all models on the HELP downward monotone subset ($\sim$ 21k examples). The right-most column of Table 4 shows that while ESIM and BERT achieve very high development accuracy on the upward data, they fail to generalize to the downward monotone test set. The proposed models generalize well and achieve better test accuracy on the downward monotone datasets.

Aggregation Decisions: The proposed model provides inference explainability by accessing natural logic's aggregation and decision paths. Figure 3 shows an example of the 2-hop dataset, together with the visualization of the intermediate aggregation decisions. From left to right, the first subfigure shows the cross-sentence attention between the premise (x-axis) and hypothesis (y-axis), where a darker color corresponds to a larger attention weight. In the second subfigure, for each word in the hypothesis (y-axis), the predicted distribution of lexical-level logical relations are shown along the x-axis. The third subfigure shows the aggregation output. For example, on the second row, the aggregation has already been performed over the first two words $\boldsymbol{b}_1$ ="*the*" and $\boldsymbol{b}_2$ ="*animals*" using their lexical relation distributions, which have been shown in the second subfigure and are, in turn, computed from the first subfigure using $\langle \tilde{\boldsymbol{b}}_1, \boldsymbol{b}_1 \rangle$ and $\langle \tilde{\boldsymbol{b}}_2, \boldsymbol{b}_2 \rangle$. Since '$\equiv$' $\bowtie$ '$\sqsubset$' = '$\sqsubset$', we can see that on the second row, a large probability mass has been put on $\sqsubset$ (i.e., *ent_f* in the figure).

We further perform quantitative analysis on the aggregation performance. We analyze the sequential aggregation. Specifically, for the 2-hop dataset in which we have access to the aggregation decisions: $\hat{\boldsymbol{z}} = \{\hat{z}_1, \hat{z}_2, ..., \hat{z}_n\}$, where $\hat{z}_j$ is the aggregation result at time step j, we evaluate the models by comparing the estimated $\hat{\boldsymbol{z}}$ with the ground truth $\boldsymbol{z}$. We use precision, recall, and F-score as our evaluation metrics. The details of how to compute them are in Appendix B.

Table 5 shows the results. Since ESIM and BERT do not produce intermediate aggregation results, they are not included in the table. The ablation analysis shows that both the memory/module component and the local relation constraints help the model to learn intermediate natural logic aggregation. We can also see that further work is desirable to improve the performance on aggregation prediction as there is still a large room to improve modeling performance on this. As part of our efforts, we have also performed component training to leverage WordNet (Miller, 1998) and ConceptNet (Speer et al., 2017) to help determine lexical relations. This approach is not particularly effective since the lexical pairs from these knowledge bases only cover a very small percentage of pairs that need to be modeled.

6 Conclusions

This paper studies end-to-end trained differentiable models that integrate natural logic with neural networks. The proposed model adapts module networks to model natural logic operations, which is

Premise: Two dogs, the gray poodle high in the air, play on the grass.
Hypothesis: The animals are lying on the bed.
Label: Contradiction

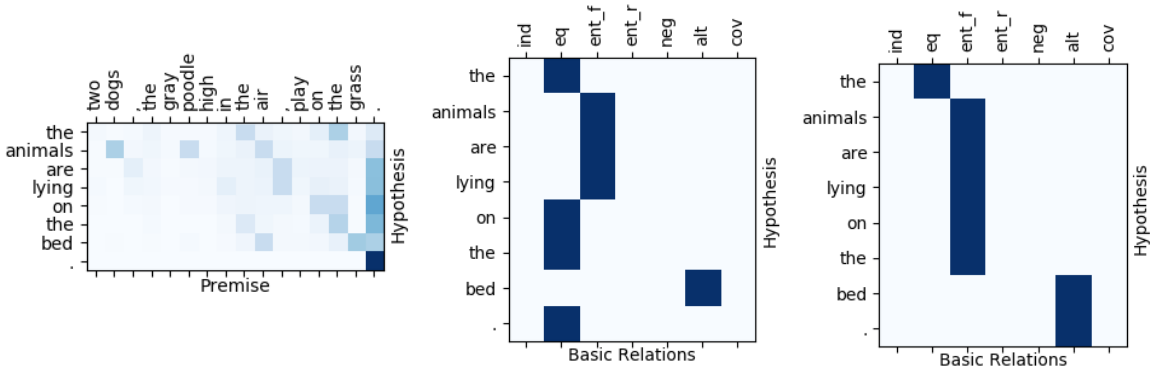

Figure 3: An example showing how the proposed model perform natural logic aggregation.

enhanced with a memory component to model contextual information. We analyze the proposed model on the monotonicity subset of Semantic Fragments, HELP, MED, and a subset of MED that are modified to include 2-hop inference. Our experiments show that the proposed framework can effectively model monotonicity-based reasoning, compared to the two baseline neural network models without built-in inductive bias for monotonicity-based reasoning. The proposed model show to be robust when transferred from upward to downward inference. We perform further analyses on the performance of the proposed model on aggregation, showing the effectiveness of the proposed subcomponents on helping achieve better intermediate aggregation performance.

7 Acknowledgement

The first, second, fourth and last author's research is supported by NSERC Discovery Grants.

References

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, United States.

Gabor Angeli and Christopher D Manning. 2014. Naturalli: Natural logic inference for common sense reasoning. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, Doha, Qatar.

Gabor Angeli, Neha Nayak, and Christopher D Manning. 2016. Combining natural logic and shallow reasoning for question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berlin, Germany.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2017a. Neural natural language inference models enhanced with external knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017b. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, Vancouver, Canada.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017c. Recurrent neural network-based sentence encoder with gated attention for natural language inference. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, Copenhagen, Denmark, September.

Qian Chen, Zhen-Hua Ling, and Xiaodan Zhu. 2018. Enhancing sentence embedding with generalized pooling. In *Proceedings of the 27nd International Conference on Computational Linguistics (COLING)*, Santa Fe, New Mexico, USA.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Proceedings of the First international conference on Machine Learning Challenges: evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*.

Luc De Raedt, Robin Manhaeve, Sebastijan Dumancic, Thomas Demeester, and Angelika Kimmig. 2019. Neuro-symbolic= neural+ logical+ probabilistic. In *NeSy'19@ IJCAI, the 14th International Workshop on Neural-Symbolic Learning and Reasoning*, Macao, China.

Thomas Demeester, Tim Rocktäschel, and Sebastian Riedel. 2016. Lifted rule injection for relation embeddings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, Texas.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, USA.

Michelangelo Diligenti, Marco Gori, and Claudio Sacca. 2017. Semantic-based regularization for learning and inference. In *Artificial Intelligence*, volume 244, pages 143–165.

I Donadello, L Serafini, and AS d'Avila Garcez. 2017. Logic tensor networks for semantic image interpretation. In *26th International Joint Conference on Artificial Intelligence (IJCAI)*, Melbourne, Australia.

Honghua Dong, Jiayuan Mao, Tian Lin, Chong Wang, Lihong Li, and Denny Zhou. 2019. Neural logic machines. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, Long Beach, California, USA.

Richard Evans and Edward Grefenstette. 2018. Learning explanatory rules from noisy data. In *Journal of Artificial Intelligence Research (JAIR)*, volume 61, pages 1–64.

Artur d'Avila Garcez, Tarek R Besold, Luc De Raedt, Peter Földiak, Pascal Hitzler, Thomas Icard, Kai-Uwe Kühnberger, Luis C Lamb, Risto Miikkulainen, and Daniel L Silver. 2015. Neural-symbolic learning and reasoning: contributions and challenges. In *2015 AAAI Spring Symposium Series*, Austin, Texas, USA.

Nitish Gupta, Kevin Lin, Dan Roth, Sameer Singh, and Matt Gardner. 2020. Neural module networks for reasoning over text. In *Proceedings of the 8nd International Conference on Learning Representations (ICLR)*, Formerly Addis Ababa, Ethiopia.

Kelvin Guu, Panupong Pasupat, Evan Zheran Liu, and Percy Liang. 2017. From language to programs: Bridging reinforcement learning and maximum marginal likelihood. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, Vancouver, Canada.

Thomas F Icard and Lawrence S Moss. 2014. Recent progress on monotonicity. In *Linguistic Issues in Language Technology*. Citeseer.

Thomas F Icard. 2012. Inclusion and exclusion in natural language. *Studia Logica*.

Adrian Iftene and Alexandra Balahur-Dobrescu, 2007. *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. Prague, Czech.

Douwe Kiela, Changhan Wang, and Kyunghyun Cho. 2018. Dynamic meta-embeddings for improved sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA.

George Lakoff. 1970. Linguistics and natural logic. *Synthese*.

Tao Li and Vivek Srikumar. 2019. Augmenting neural networks with first-order logic. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, Austin, Texas, United States.

Bill MacCartney and Christopher D Manning. 2008. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, Manchester, UK.

Bill MacCartney and Christopher D Manning. 2009. An extended model of natural logic. In *Proceedings of the 8th international conference on computational semantics (IWCS)*, Stroudsburg, United States.

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

Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. 2018. Deepproblog: Neural probabilistic logic programming. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*, Montreal, Canada.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2019. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, New Orleans, USA.

George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

Sewon Min, Danqi Chen, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. A discrete hard em approach for weakly supervised question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Hong Kong, China.

Pasquale Minervini, Matko Bošnjak, Tim Rocktäschel, Sebastian Riedel, and Edward Grefenstette. 2020. Differentiable reasoning on large knowledge bases and natural language. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*, New York, USA.

Rowan Nairn, Cleo Condoravdi, and Lauri Karttunen. 2006. Computing relative polarity for textual inference. In *Proceedings of the 5th international workshop on inference in computational semantics*, Buxton, England.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, Doha, Qatar.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2227–2237, New Orleans, USA.

Kyle Richardson, Hai Hu, Lawrence S Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, New York, USA.

Tim Rocktäschel and Sebastian Riedel. 2017. End-to-end differentiable proving. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, Long Beach, USA.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI)*, San Francisco, California USA.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in Neural Information Processing Systems*, Montréal Canada.

Aarne Talman, Anssi Yli-Jyrä, and Jörg Tiedemann. 2018. Natural language inference with hierarchical bilstm max pooling architecture. *arXiv preprint arXiv:1808.08762*.

Víctor Manuel Sánchez Valencia. 1991. *Studies on natural logic and categorial grammar*. Universiteit van Amsterdam.

Johan Van Benthem. 1986. *Essays in logical semantics*. Springer.

Johan van Benthem. 1988. The semantics of variety in categorial grammar. *Categorial grammar*.

Johan Van Benthem. 1995. *Language in Action: categories, lambdas and dynamic logic*. MIT Press.

Leon Weber, Pasquale Minervini, Jannes Münchmeyer, Ulf Leser, and Tim Rocktäschel. 2019. Nlprolog: Reasoning with weak unification for question answering in natural language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, Austin, Texas, United States.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *arXiv preprint arXiv:1410.3916*.

Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Van den Broeck. 2018. A semantic loss function for deep learning with symbolic knowledge. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, Stockholm, Sweden.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019a. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Austin, Texas, United States.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019b. Help: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM)*, Minneapolis, Minnesota, USA.

Fan Yang, Zhilin Yang, and William W Cohen. 2017. Differentiable learning of logical rules for knowledge base reasoning. In *Advances in Neural Information Processing Systems*, Long Beach, USA.

Xiaoyu Yang, Xiaodan Zhu, Huasha Zhao, Qiong Zhang, and Yufei Feng. 2019. Enhancing unsupervised pretraining with external knowledge for natural language inference. In *Canadian Conference on Artificial Intelligence*, pages 413–419. Springer.

Deunsol Yoon, Dongbok Lee, and SangKeun Lee. 2018. Dynamic self-attention : Computing attention over words dynamically for sentence embedding. *arXiv preprint arXiv: 1808.07383*.

A The 2-hop Dataset

Figure 4 shows an example of the 2-hop dataset. The premise and hypothesis differ by two edits of word/phrase insertion, deletion, or substitution. The dataset provides the ground truth of aggregation at each time step (the *equivalence* relation is the default relation and is hence not included in the "Ground truth of aggregation" section) and the word locations/indices associated with each edit. The 2-hop dataset is developed with the following three steps:

Premise: Some delegates finished the survey on time.
Hypothesis: Some individuals finished the survey.
Label: Entailment

Edit 1: delegates → individuals **Type**: hypernym **Relation**: forward_entailment
Location: premise: {1} hypothesis: {1}

Edit 2: on time → [ ] **Type**: delete **Relation**: forward_entailment
Location: premise: {5, 6} hypothesis: { }

Ground truth of aggregation:
Position {1}: forward entailment
Position {4}: forward entailment

Figure 4: An example of the 2-hop dataset.

Identifying MED Relations: Since most sentence pairs in the MED dataset are only different by one word/phrase edit; i.e., the premise and the hypothesis differs by one word/phrase, it is easy to determine location of the insertion, deletion, or replacement. For insertion and deletion, we follow (Angeli and Manning, 2014) and treat the relation as *reverse entailment* ($\sqsupset$) and *forward entailment* ($\sqsubset$), respectively. We set aside the replacement samples since we can not determine their relations without human labeling. To ensure the identified natural logic relations are correct, we compare the labels provided in MED with labels determined by MacCartney's natural logic theory and remove samples in which labels do not agree, yielding roughly 1.1K sentence pairs.

Adding One More Hop of Relations: We ask human annotators to replace a noun either in the premise or the hypothesis with another word. The relation between the substituted and substituting word are one of $\{\equiv, \sqsubset, \sqsupset, |, \#\}$. Annotators have access to WordNet that can help suggest substituting words (e.g., hypernyms or hyponyms). Meanwhile, we require that the candidate words to be replaced are not children or parents of any previously identified differences over the parsing tree. This replacement operation yields 5,858 sentence pairs, and the premise and the hypothesis of each example now differ by two edits.

Determining Labels: We apply projection operation and natural logic aggregation according to (MacCartney and Manning, 2009) to determine the 3-way natural language inference labels for the generated 2-hop sentence pairs. We also record the ground-truth relations of each hop of aggregation output. We manually assess the data quality on 300 sentence pairs (100 for each category). We find that on average 3% of the samples have either incorrect labels or wrong intermediate aggregation output (4% in category *Entailment*, 4% in category *Neutral* and 1% in *Contradiction*). Those mistakes are mainly produced by incorrect parser-identified polarity.

B Aggregation Evaluation Metrics

We evaluate the intermediate aggregations of the proposed model with the precision, recall, and F1 score. Precision is the number of correctly performed aggregations, divided by the total number of aggregations performed by a model. Recall is the number of correctly performed aggregations, divided by the total number of aggregations presented in the ground-truth annotation. Note that we only consider aggregations at time step t when $\hat{z}_t \neq \hat{z}_{t-1}$. Since by default the starting state $\hat{z}_0 =$ '$\equiv$', so if $\hat{z}_1 =$ '$\equiv$', we do not count this degenerate case.