Neural Fuzzy Logic Reasoning for Natural Language Inference

by

Zijun Wu

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science
University of Alberta

Abstract

Natural language inference, also known as NLI, aims to determine the logical relationship between two sentences, such as Entailment, Contradiction, and Neutral. NLI is important to natural language processing, because it involves logical reasoning and is a key problem in artificial intelligence. In recent years, deep learning models have become a prevailing approach to NLI. Those approaches can achieve high performance, but lack interpretability and explainability.

In this work, we propose an Explainable Phrasal Reasoning (EPR) approach to address the explainability for NLI by weakly supervised logical reasoning. The system includes three main components. It first detects phrases as the semantic unit and aligns corresponding phrases. Then, it predicts the NLI label for the aligned phrases, and induces the sentence label by fuzzy logic formulas. Our EPR is almost everywhere differentiable and thus the system can be trained end-to-end in a weakly supervised manner. We annotated a corpus and developed a set of metrics to evaluate phrasal reasoning. Results show that our EPR yields much more meaningful explanations in terms of F scores than previous studies. To the best of our knowledge, we are the first to develop a weakly supervised phrasal reasoning model for the NLI task.

"Murphy's Law doesn't mean that something bad will happen.

It means that whatever can happen will happen"

- Interstellar

For my parents, who support me unconditionally.

And in memory of my grandmother.

Acknowledgments

I would like to thank my supervisor Dr. Lili Mou for his support in the past two years. He inspired me to make progress on this work, because he was always excited when we had discussions about it, and felt what I was doing is interesting and important. He gave me very insightful suggestions on the experimental design, especially the evaluation metric for reasoning. He also edited and proofread my thesis. I couldn't finish this thesis without his effort on it.

A preliminary manuscript of the thesis work is available at https://arxiv.org/pdf/2109.08927.pdf. As the first author of the project, I devised the methodology and led the entire research process. But this work is not my own alone. Dr. Mou contributed to this work with no doubt. I very appreciate the contributions from the second and third authors of this project. They are Atharva Naik from Indian Institute of Technology Kharagpur, and Zixuan Zhang from University of Alberta. Atharva came to our research group for the summer internship, and helped me for developing the web annotation interface in my experiment, with his professional web development skill. Zixuan is my friend and collaborator in our group; we brainstormed every time we were on the trip to snowboarding, and he contributed the idea of heuristic phrase alignment to my methodology.

I am thankful to my parents. Their support encourages me to move on without hesitation. I also want to give thanks to Prof. Davood Rafiei, who guided me for completing my first NLP project in my Master's course. Finally, I want to thank the University of Alberta for funding me continuously through providing research and teaching assistantship during the past two years.

The research is supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) under grant No. RGPIN2020-04465, the Amii Fellow Program, the Canada CIFAR AI Chair Program, a UAHJIC project, a donation from DeepMind, and Compute Canada (www.computecanada.ca).

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Deep learning with neural networks have become the most commonly used method in many real-world applications, including image classification [28, 18], information retrieval [26], machine translation [2], and machine reading comprehension [22, 12]. Thanks to the recent advance in large neural modeld and large annotated datasets, deep learning models are so powerful that they have already beaten human performance in certain tasks. Take machine reading comprehension (MRC) task as an example. Neural MRC models based on BERT [12] are able to exceed human performance on the SQuAD MRC dataset [49].

Besides the high performance that deep neural networks can achieve, the interpretability of neural models is another important research [15], as the predictions from such neural models are "unpredictable", i.e., even the developers cannot explain why the model makes a specific decision. Thus, lack of interpretability of neural model may lead to a variety of issues related to security and bias [53]. Recently, the European Commission proposed a European Union regulatory framework on artificial intelligence[1]. It aims to support the usage of AI while lowering the risks, by considering interpretation as an essential component of AI systems.

In this thesis, we focus on the interpretation of the Natural Language Infer-

[1] https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf

ence (NLI) task. NLI aims to determine the logical relationship between two sentences (called a *premise* and a *hypothesis*), and the target labels include `Entailment`, `Contradiction`, and `Neutral` [5, 39]. Table 1.1 gives an example, where the hypothesis contradicts the premise. NLI is important to natural language processing, because it involves logical reasoning and is a key problem in artificial intelligence. Previous work shows that NLI can be used in various downstream tasks, such as information retrieval [26] and text summarization [34].

In recent years, deep learning has become a prevailing approach to NLI [5, 43, 56, 62]. Especially, pretrained language models with the Transformer architecture [54] achieve state-of-the-art performance for the NLI task [48, 65]. However, such deep learning models are black-box machinery and lack interpretability. In real applications, it is important to understand how these models make decisions [51].

Recently, several studies have addressed the explainability of NLI models. For example, Camburu et al. [7] generate a textual explanation by sequence-to-sequence supervised learning, in addition to NLI classification; such an approach is multi-task learning of text classification and generation, which requires additional human annotations but does not perform reasoning itself. MacCartney et al. [36] propose a scoring model for aligning relative phrases; Parikh et al. [45] and Jiang et al. [24] propose to obtain alignment by attention mechanisms; however, they only provide correlation information, instead of logical reasoning. Other research incorporates upward and downward monotonicity entailment reasoning for NLI [21, 9], but these approaches are based on hand-crafted rules (e.g., *every* downward entailing *some*) and are restricted to `Entailment` only; they cannot handle `Contradiction` or `Neutral`. None of previous studies can perform expressible logical reasoning for the NLI task. Nor have they quantitatively evaluated reasoning performance.

Input:
Premise: Several men helping each other pull in a fishing net.
Hypothesis: There is one man holding the net.
Sentence-Level Prediction:
[ ] Entailment [✓] Contradiction [ ] Neutral
Phrase-Level Reasoning:
`Entailment`: *pull in a fishing net* VS *holding the net*
`Contradiction`: *several men* VS *one man*
`Neutral`: (None)
`Unaligned phrase(s)`: *helping each other*

Table 1.1: The natural language inference (NLI) task and a desired reasoning mechanism.

1.2 Explainability For NLI

In this thesis, we address the explainability for NLI by weakly supervised logical reasoning[2]. Intuitively, an NLI system with an explainable reasoning mechanism should be equipped with the following functionalities:

1. The system should be able to detect corresponding phrases and tell their logical relationship, e.g., *several men* contradicting *one man*, but *pull in a fishing net* entailing *holding the net* in Table 1.1.
2. The system should be able to induce sentence labels from phrase-level reasoning. In the example, the two sentences are contradictory because there exists one contradictory phrase pair.
3. More importantly, such reasoning should be trained in a weakly supervised manner, i.e., the phrase-level predictions are trained from sentence labels only. Otherwise, the reasoning mechanism degrades to multi-task learning, which requires massive fine-grained human annotations.

To this end, we propose an Explainable Phrasal Reasoning (EPR) approach to

[2]Our phrase-level logical reasoning for the NLI task is restricted to three NLI labels: `Entailment`, `contradiction`, `neutral`

the NLI task. Our model uses heuristics to obtain phrases as semantic units and aligns corresponding phrases by embedding similarity. Then, we predict the NLI labels (namely, `Entailment`, `Contradiction`, and `Neutral`) for the aligned phrases. Finally, the sentence prediction is induced from the phrasal NLI labels. For example, two sentences are contradictory if there exists a contradictory phrase pair. Such reasoning is accomplished in a fuzzy logic manner [63, 64]; thus, our model is differentiable and the phrasal reasoning component can be trained with weak supervision of sentence NLI labels. In this way, our EPR approach satisfies all the desired properties mentioned above.

In our experiments, we developed a comprehensive methodology (data annotation and evaluation metrics) to quantitatively evaluate reasoning performance, which has not been accomplished in previous work. Since no previous work can provide expressible phrasal logic reasoning, we tried our best to extend previous studies and obtain plausible baseline models. Results show that our EPR yields much more meaningful explanations in terms of F scores against human annotation.

1.3 Thesis Statement

The objective of this thesis is to improve the interpretability of neural model for the NLI task. We argue that neural model can provide phrase-level reasoning for its sentence-level prediction. Our thesis contributions are summarized as follows:

1. We formulate a phrasal reasoning task for natural language inference (NLI), addressing the importance of interpretability of neural text understanding models.
2. We propose EPR that induces sentence-level NLI labels from explainable phrasal reasoning by neural fuzzy logic. EPR is able to perform reasoning in a weakly supervised way.
3. We created an annotated corpus and a set of metrics to evaluate phrasal reasoning. We release the code and annotated data for future studies.

To the best of our knowledge, we are the first to develop a weakly supervised

phrasal reasoning model for the NLI task.

1.4 Thesis Organization

In this chapter, we introduce why explainability is important to neural network. We also address the problem of the explainability for NLI models. As an overview for this thesis, we show that our propsed EPR model has the best reasoning ability among all baseline models. It also has impressive explainability that can even detect errors in the dataset.

The remainder of the thesis contains four chapters.

In Chapter 2, we introduce the background knowledge on language modeling as well as the related work on addressing NLI's explainability.

In Chapter 3, we describe the proposed methodology regarding reasoning for NLI. Our EPR model utilizes differentiable fuzzy logic so that the reasoning is trained in a weakly supervised manner.

In Chapter 4, we conduct comprehensive experiments that include quantitative performance on two NLI datasets and a case study.

In Chapter 5, we conclude the entire thesis, and show the limitation of EPR that can be addressed in the future.

Chapter 2

Background and Related Work

Natural language processing (NLP) is an important component of Artificial Intelligence (AI). The goal of NLP is to make machine to understand and generate human language, which involve natural language understanding (NLU) and natural language generation (NLG). Our work focuses on NLU, as the reasoning problem is to endow machine the ability to explain the process of how it understands natural language.

In this chapter, we introduce the background knowledge for our work. Section 2.1 describes recent studies on Natural Language Understanding including language modeling and its applications. Section 2.2 introduces Natural Language Inference (NLI), a specific task of NLU. Our work studies the reasoning problem of NLI. In Section 2.2, we introduce recent studies for NLI and its reasoning. Our reasoning mechanism utilizes old-school fuzzy logic and prevailing deep learning methods, and we introduce neuro-symbolic approaches and fuzzy logic in Sections 2.3 and **??**, respectively.

2.1 Natural Language Understanding

2.1.1 Language Modeling

A language model determines the probability over sequences of words, which models grammatical correctness and fluency of word sequences. Language model is also able to assign a probability to the next word given a sequence of words. Given a word sequence "How are", a language model estimates the probability of all possible words

(i.e., all words in a pre-defined English vocabulary) and the word "you" may have the highest probability to be the next word, because "How are you" is the most commonly used English phrase in our daily life.

Language modeling is the foundation of many NLP tasks regarding both NLU (e.g., sentiment analysis, text classification) and NLG (e.g., summarization, translation), and is a core component of modern NLP systems [25].

Formally, a language model yields the probability of a sequence of words $w_1, \cdots, w_T$ with length T, denoted by

$$P(w_1, \cdots, w_T) \tag{2.1}$$

The probability of each word w_t in the sentence is conditioned on the preceding word sequence $w_1, \cdots, w_{t-1}$ that are generally called *context*.

$$P(w_1, \cdots, w_t) = P(w_t \mid w_1, \cdots, w_{t-1}) P(w_1, \cdots, w_{t-1}) \tag{2.2}$$

Thus, we can factorize the sequence probability from Equation 2.1 into

$$P(w_1, \cdots, w_T) = P(w_1) P(w_2 \mid w_1) \cdots P(w_T \mid w_1, \cdots, w_{T-1}) \tag{2.3}$$

However, the language model described above has an extremely large number of parameters to learn if we directly parametrize each probability as a multinomial distribution. Thus, approximation methods are needed. The N-gram model is an approximation that was widely used before the Neural Net Language Model (NNLM). In the N-gram model, we assume that the probability of the current word only depends on the previous $n-1$ word:

$$P(w_t \mid w_1, \cdots, w_{t-1}) \approx P(w_t \mid w_{t-n}, \cdots, w_{t-1}) \tag{2.4}$$

Although the N-gram model can reduce the size of parameters, it still has a fundamental problem which is the *curse of dimensionality* [4]. Because of the discreteness nature of N-gram, the problem becomes more severe when we want to model the joint distribution between many discrete words in a sentence [4]. For example, if we want

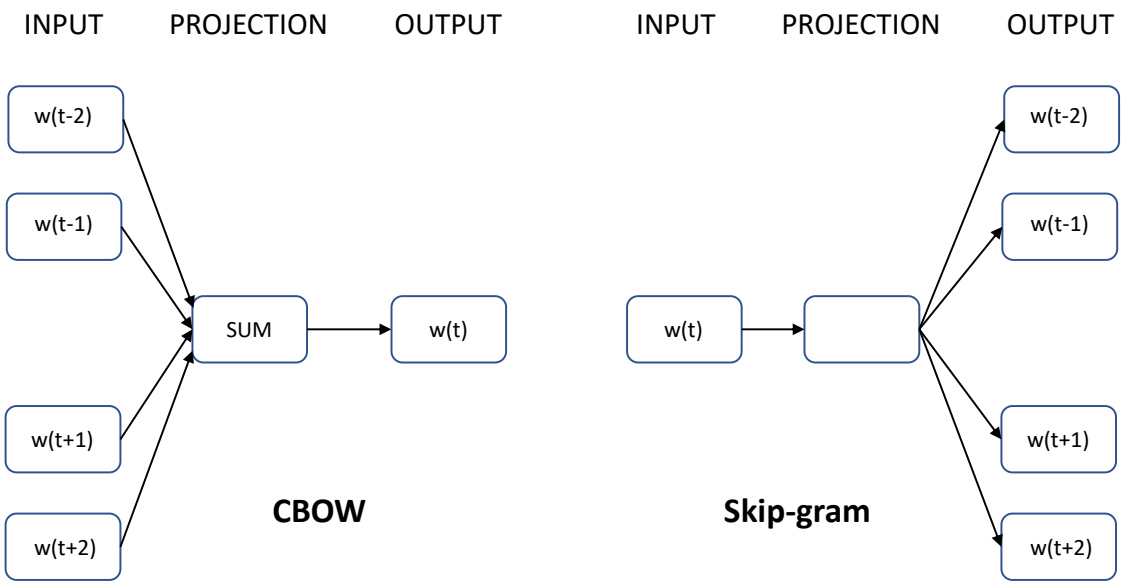

Figure 2.1: The model structures of CBOW and Skip-gram. The figure is adapted from [41].

to model a 10-gram language model with a vocabulary of size 10000, there are 10^{50} free parameters.

The neural net language model (NNLM) is introduced to address the problem because of its property of continuity. Bengio et al. [4] propose the first NNLM by learning a distributed representation for words, which is also known as a *word embedding*. Each word embedding is a feature vector of real numbers, and it can further reduce the size of parameters because word embeddings are usually dense and low-dimensional. CBOW and Skip-gram [41] further improve the word representation by self-supervised learning on the Google News corpus. As shown in Figure 2.1, CBOW is trained to predict the current word based on the surrounding words, whereas Skip-gram predicts the surrounding words given the current word. It shows some intriguing properties. For example, both "China" and "France" are countries, and the similarity between their word embeddings is close. Moreover, word embedding can show semantic relationships between words with such an equation: $China - Beijing \approx France - Paris$, since Beijing is the capital of China and Paris is the capital of France [41].

However, word embeddings from CBOW or Skip-gram only contain the local se-

mantics without considering the interaction with other words that appeared in the same sentence or document. For example, the word "apple" can refer to a fruit apple or a company called Apple, where its semantics should be based on the context surrounding it. Using the Recurrent Neural Network (RNN) [19, 10] is a natural way to get the contextual word embeddings, since the same parameters are shared across all time steps. Language model that combines RNN and the pre-trained word embeddings is a common practice before Transformer [54] is proposed.

2.1.2 Transformer

The Transformer [54] is a predominant neural model architecture that is widely used in many NLP tasks. It is originally proposed as a sequence-to-sequence model with an encoder and a decoder, which is suitable for natural language generation tasks such as machine translation. Specifically, a sequence of words is first encoded by the encoder into a list of contextual word embeddings. The decoder then decodes the embeddings into a probability distribution of the next possible words, which is conditioned on the previously predicted words.

The encoder and decoder of the Transformer consist of a stack of N identical blocks. Each encoder block has two key components, namely, a self-attention module and a feed-forward neural network. A residual connection [18] is employed around each module for building the deeply stacked blocks, followed by a layer normalization module [1]. Compared with the encoder block, a decoder block has an additional cross-attention module between the self-attention module and the feed-forward neural network. The architecture of Transformer is shown in Figure 2.2. After applying a linear and a softmax layer, the model can output a probability distribution of the next word over the whole vocabulary.

What makes Transformer a successful language model is mainly because it utilizes attention mechanisms. The attention used in the Transformer architecture is particularly called "Scaled Dot-Product Attention". The input consists of the matrix

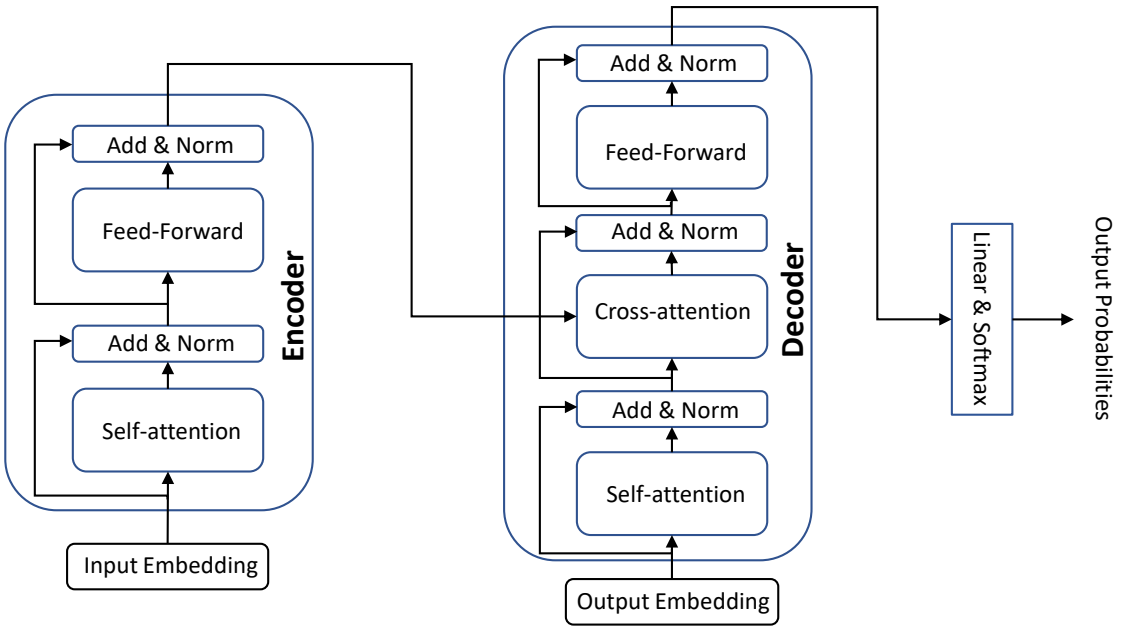

Figure 2.2: The model structures of Transformer. The figure is adapted from [54].

representations of queries $Q \in \mathbb{R}^{M \times D_k}$, keys $K \in \mathbb{R}^{N \times D_k}$, and values $V \in \mathbb{R}^{N \times D_v}$, where M and N are the lengths of queries and keys (or values); D_k and D_v denote the dimensions of queries (or keys) and values. Formally, the attention is given by

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{D_k}}\right)V \tag{2.5}$$

The dot product of queries and keys is scaled by $\frac{1}{\sqrt{D_k}}$ because it can alleviate the gradient vanishing issue caused by the softmax operation.

Pre-training based on Transformer provides the opportunity of not only fine-tuning the model in a downstream task, but also using it in an off-the-shelf manner. Generally, the Transformer architecture can be used three different modes:

- Encoder–decoder mode: such as BART [30], which is typically useful for the sequence-to-sequence generation tasks;
- Encoder-only mode: such as BERT [12], which outputs representations for the input sequence and can be used for classification tasks with a subsequent linear layer and softmax function; and

- Decoder-only mode: such as GPT [6], which is typically used for sequence generation such as language modeling.

In this thesis, we utilize Sentence-Transformer [50], which uses a pre-trained encoder-only Transformer as the backbone, and is further fine-tuned using paraphrase datasets to learn high-quality sentence-level embeddings.

2.1.3 Sentence Embedding

Sentence embedding represents the text semantics as real-valued vectors for a sentence, which is usually used for the tasks such as information retrieval and semantic similar comparison. Similar to word embeddings, the goal is to embed sentences into a vector space so that semantically similar sentences are close in the vector space. However, the difference is that sentences may vary in length, whereas a word is a single token. To obtain the embedding of a sentence, an intuitive approach is to average the word embeddings in the sentence. Although such a method of is efficient, the performance may be unsatisfactory.

The most common approach is to encode the sentences using the pretrained BERT model [12], and treat the output of the first token (the [CLS] token) as the embedding of the input sentences. However, the performance of this approach is even worse than averaging word embeddings [50]. This is because the [CLS] token is used for Next Sentence Prediction (NSP) task in pre-training, which is learned to describe the temporal relation between sentences in vector space. So it is not suitable to use it to represent the semantics of a sentence directly.

Reimers and Gurevych [50] propose Sentence-Transformer, and improve the quality of sentence embeddings drastically. Specifically, they fine-tuned the pre-trained Transformer model with the paraphrase or NLI datasets, based on the [CLS] token or mean pooling of the last layer. As shown in Figure 2.3, two sentences from the training data are fed into two BERT models with tied weights; after fine-tuning, only one model needs be retained. The outputs u and v are the corresponding sentence

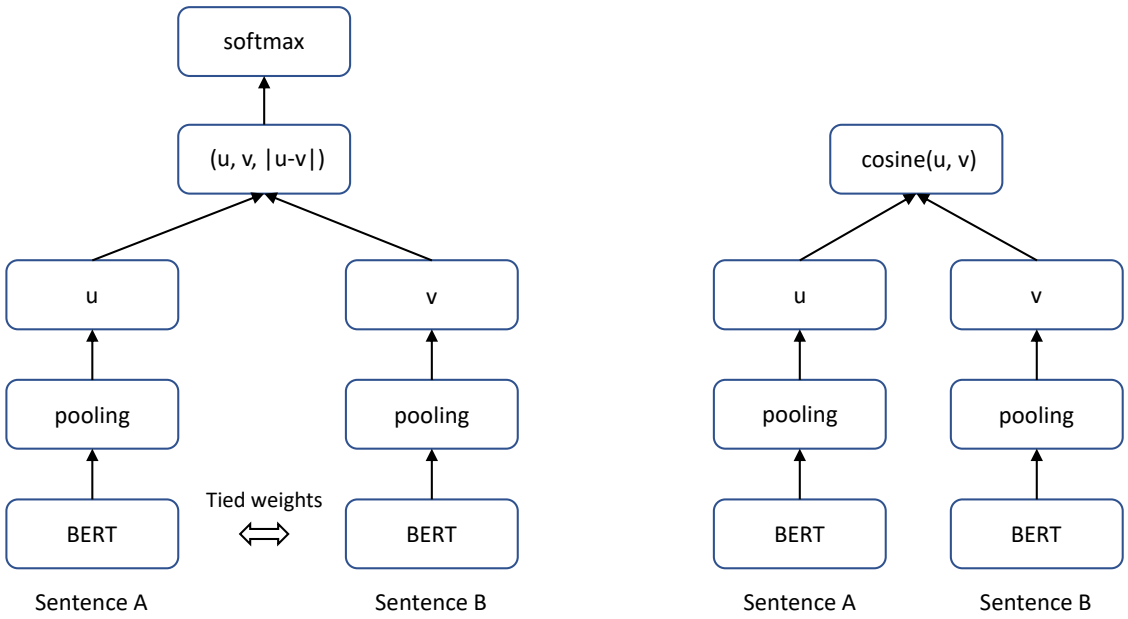

Figure 2.3: The architecture of SBERT, where BERT is used as the backbone Transformer model. The left figure is SBERT during fine-tuning, where two BERT models have tied weights (also known as the Siamese network structure). The right figure is the SBERT applied to similarity modeling. The figure is adapted from [50].

embeddings. They are then combined with their element-wise difference $|u - v|$ to form a single vector. When fine-tuning on paraphrase datasets, a feed-forward neural net followed by a softmax function performs two-way classification, whereas for NLI datasets, a three-way classifier is needed.

The fine-tuned BERT model is called the SBERT model, which can be used off the shelf to provide sentence embeddings. The output sentence embeddings are computed by the cosine function to compare the similarity of the corresponding two sentences. We utilize SBERT at the phrase level to match the extracted phrases with similar semantics. It is also used to provide phrasal embeddings that predict the NLI relations for the matched phrase pairs.

2.2 Natural Language Inference Task

Natural Language Inference (NLI) task, also known as Recognizing Textual Entailment (RTE), aims to determine whether the meaning of one sentence (hypothesis) can

be inferred from another (premise) [5]. A variety of techniques have been addressed on NLI, including symbolic logic and neural networks. Although symbolic logic is able to provide explicit reasoning for NLI, the ultimate performance is oftentimes lower. This is because the knowledge coverage is restricted by manually designed logical rules, so that the model is not possible to obtain a powerful generalization ability. On the other hand, deep learning models benefit from crowd-sourced NLI datasets, such as SNLI [5] and MultiNLI [57], constantly refreshing the state-of-the-art performance. However, one drawback is that such a model is back-box machinery and cannot provide any explainability for its NLI prediction. Moreover, recent research [47] shows that deep learning models tend to utilize dataset bias for prediction.

There are several studies addressing reasoning in NLI. MacCartney and Manning [37] propose seven natural logic relations in addition to `Entailment`, `Contradiction`, and `Neutral`, shown in Table 2.1. MacCartney and Manning [38] also distinguish upward entailment (*every mammal* upward entailing *some mammal*) and downward entailment (*every mammal* downward entailing *every dog*) as different categories. Manually designed lexicons and rules are used to interpret `Entailment` in such a finer-grained manner [21, 9]. Feng et al. [16] apply such natural logic to NLI reasoning in the word level; however, our experiments will show that their word-level treatment is not an appropriate granularity, and that they fail to achieve meaningful reasoning performance.

The above reasoning schema focuses more on the quantifiers of first-order logic (FOL) [3]. However, the SNLI dataset [5] only contains less than 5% samples with explicit quantifiers, and the seven-category schema complicates reasoning in the weakly supervised setting. Instead, we adopt three-category NLI labels following the SNLI dataset. Our focus is entity-based reasoning, and the treatment of quantifiers is absorbed into phrases.

We also notice that previous work lacks explicit evaluation of the reasoning performance for NLI. For example, the SNLI dataset only provides sentence-level labels.

Relation	Name	Example	Mapping
x $\equiv$ y	equivalence	dad $\equiv$ father	entailment
x $\sqsubset$ y	forward entailment	puppy $\sqsubset$ dog	entailment
x $\sqsupset$ y	reverse entailment	cat $\sqsupset$ kitten	neutral
x ^ y	negation	human ^ nonhuman	contradiction
x | y	alternation	cat | dog	contradiction
x $\smile$ y	cover	animal $\smile$ nonhuman	neutral
x # y	independence	cat # sleep	neutral

Table 2.1: Seven natural logical relations [37], and the mapping to three-category labels.

The HELP [60] and MED [59] datasets concern monotonicity inference problems, where the label is also at the sentence level; they only consider `Entailment`, ignoring `Contradiction` and `Neutral`. Thus, we propose a comprehensive framework for the evaluation of NLI reasoning.

2.3 Neuro-Symbolic Approaches

In recent years, neuro-symbolic approaches have attracted increasing interest in the AI and NLP communities for explaining and interpreting deep learning models. Typically, these approaches are trained by reinforcement learning or its relaxation, such as attention and Gumbel-softmax [23], to reason about certain latent structures in a downstream task.

For example, Lei et al. [29] and Liu et al. [33, 32] extract key phrases for a text classification task. The key phrase extraction is learned jointly with the classification to provide the meaningful rationale for the classification prediction, yet it is only supervised by the classification task. Lu et al. [35] extract entities and relations for document understanding. Specifically, the proposed OONP model reads the document and parses the entities into the object-oriented data structure. This parsing process can either be trained with supervised learning, reinforcement learning, or a

combination of both. Liang et al. [31] and Mou et al. [42] perform SQL-like execution based on input text for semantic parsing. They both train a symbolic executor with reinforcement learning, combining with the neural network to understand the input text. Xiong et al. [58] use a policy-based agent to hop over a knowledge graph for reasoning the relationships between entities. In this work, we address logical reasoning for the NLI task, which is not tackled in previous neuro-symbolic studies.

Mahabadi et al. [40] apply fuzzy logic formulas to replace multi-layer perceptrons for NLI. In this way, they manage to reduce the number of model parameters, but their performance is lower. Also, they are unable to provide expressive reasoning because their fuzzy logic works on sentence features. Our work is inspired by [40], but we propose to apply fuzzy logic to the detected and aligned phrases, and are able to provide reasoning in a symbolic (i.e., expressive) way. We also develop our own fuzzy logic formulas, which are different from [40].

2.4 Chapter Summary

In this chapter, we introduced the background knowledge of natural language understanding. This includes N-gram approximation for statistical language model, word embeddings, the widely used Transformer architecture which is capable of encoding and decoding natural language, and the pre-trained Sentence-Transformer for providing sentence embeddings. We then introduced the task of Natural Language Inference, with a focus on reasoning for NLI. We finally introduced relative work regarding neuro-symbolic methods and fuzzy logic.

In the next chapter, we will present our neuro-symbolic EPR approach for NLI reasoning, which combines systems from two worlds: a Transformer neural model and fuzzy logic.

Chapter 3

The EPR Approach

In this chapter, we will describe our EPR approach in detail, which is shown in Figure 3.1. It has three main components: phrase detection and alignment, phrasal NLI prediction, and sentence label induction. The first three sections describe each component in detail. In the last section, we will explain how the EPR model is trained, predicts at sentence-level and performs reasoning at phrase-level.

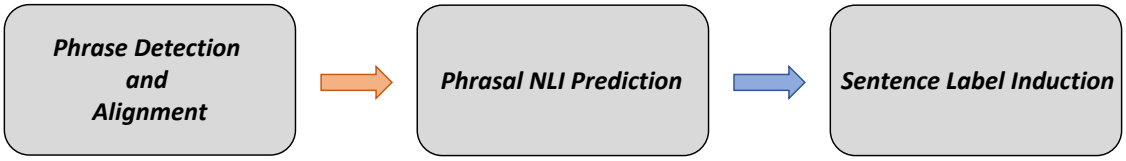

Figure 3.1: An overview of our Explainable Phrasal Reasoning (EPR) model.

3.1 Phrase Detection and Alignment

In NLI, a data point consists of two sentences, a premise and a hypothesis. We first extract content phrases from both input sentences. Compared with the word level [45, 16], a phrase presents a meaningful semantic unit, which is important to logical reasoning.

Specifically, we use SpaCy [20] to obtain the part-of-speech (POS) tag[1] of every word, and extract minimal noun phrases. In other words, no other noun phrase is

[1]See definitions in https://spacy.io/usage/linguistic-features

nested within the extracted noun phrases.

If a noun phrase follows a preposition (with a fine-grained POS tag being `IN`), it becomes a prepositional phrase. In addition, we extract verbs by the POS tag `VERB`. A verb phrase may involve a particle with the fine-grained POS tag being `RP`. Then, we treat it as a verb phrase (e.g., *show off*). In order to handle negation, we extract the pattern `AUX not VERB [RP]` as a verb phrase (e.g., *could not help*). This, however, only counts less than 1% in the dataset, and does not affect our model much.

After the above phrases are extracted, we process the remaining words based on Universal POS tags [46], which is shown in Table 3.1. Specifically, we treat remaining content words (open class words) as individual phrases. Finally, the remaining non-content words (in the categories of closed words and others) are discarded (e.g., "there is"). This is appropriate, because they do not represent meaningful semantics or play a role in reasoning. Empirically, our rule-based approach works well for the NLI dataset. Our logical reasoning is at the granularity of the extracted phrases.

We align corresponding phrases in the two sentences based on the cosine similarity. Let $\mathrm{P} = (\mathrm{p}_1, \cdots, \mathrm{p}_M)$ and $\mathrm{H} = (\mathrm{h}_1, \cdots, \mathrm{h}_N)$ be the premise and hypothesis, respectively, where p_m and h_n are extracted phrases. We apply Sentence-BERT [50] to each individual phrase and obtain the local phrase embeddings by $\boldsymbol{p}_m^{(L)} = \mathrm{SBERT}(\mathrm{p}_m)$ and $\boldsymbol{h}_n^{(L)} = \mathrm{SBERT}(\mathrm{h}_n)$. We also apply Sentence-BERT to the entire premise and hypothesis sentences to obtain the global phrase embeddings $\boldsymbol{p}_m^{(G)}$ and $\boldsymbol{h}_n^{(G)}$ by mean-pooling the features of the time steps corresponding to the words in the phrase. Their similarity is given by

$$\mathrm{sim}(\mathrm{p}_m, \mathrm{h}_n) = \gamma \cos(\boldsymbol{p}_m^{(G)}, \boldsymbol{h}_n^{(G)}) + (1 - \gamma) \cos(\boldsymbol{p}_m^{(L)}, \boldsymbol{h}_n^{(L)}) \tag{3.1}$$

where γ is the hyper-parameter to balance the lexical and contextual meaning of a phrase. It is noted that Sentence-BERT is fine-tuned on paraphrase datasets, and thus is suitable for similarity matching.

We obtain phrase alignment between the premise and hypothesis in a heuristic

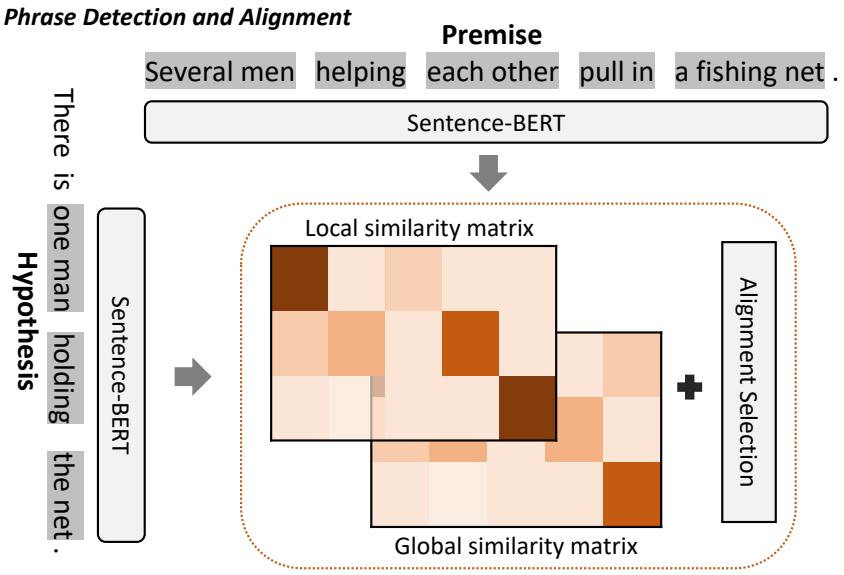

Figure 3.2: Phrase detection and alignment.

Open class words			Closed class words			Others		
POS	Definition	Example	POS	Definition	Example	POS	Definition	Example
ADJ	adjective	wonderful	ADP	adposition	during	PUNCT	punctuation	.
ADV	adverb	well	AUX	auxiliary	can	SYM	symbol	:)
INTJ	interjection	hello	CCONJ	coordinating conjunction	and	X	other	asdf
NOUN	noun	paper	DET	determiner	the			
PROPN	proper noun	John	NUM	numeral	2022			
VERB	verb	write	PART	particle	's			
			PRON	pronoun	he			
			SCONJ	subordinating conjunction	while			

Table 3.1: Universal POS tags. The table is adapted from [46].

way. For every phrase p_m in the premise, we look for the most similar phrase h_n from the hypothesis by

$$n = \mathrm{argmax}_{n'} \, \mathrm{sim}(\boldsymbol{p}_m, \boldsymbol{h}_{n'}) \tag{3.2}$$

Likewise, for every phrase h_n in the hypothesis, we look for the most similar phrase p_m from the premise. A phrase pair $(\mathrm{p}_m, \mathrm{h}_n)$ is considered to be aligned if h_n is selected as the closest phrase to p_m, and p_m is the closest to h_n. In this way, we can ensure the quality of phrase alignment, and leave other phrases unaligned (which are common in the NLI task). The process is illustrated in Figure 3.2.

3.2 Phrasal NLI Prediction

In this part, our model predicts the logical relationship of an aligned phrase pair $(\boldsymbol{p}, \boldsymbol{h})$ among three target labels: `Entailment`, `Contradiction`, and `Neutral`, as shown in Figure 3.3. While previous work [16] identifies finer-grained labels for NLI, we do not follow their categorization, because it complicates the reasoning process and makes weakly supervised training more difficult. Instead, we adopt three-label categorization for phrases, which is also consistent with sentence NLI labels.

We represent a phrase, say, p in the premise, by a vector embedding, and we consider two types of features: a local feature $\boldsymbol{p}^{(L)}$ and a global feature $\boldsymbol{p}^{(G)}$, reused from the phrase alignment component. They are concatenated as the phrase representation $\boldsymbol{p} = [\boldsymbol{p}^{(L)}; \boldsymbol{p}^{(G)}]$. Likewise, the phrase representation for a hypothesis phrase $\boldsymbol{h}$ is obtained in a similar way. Intuitively, local features force the model to perform reasoning in a serious manner, but global features are important to sentence-level prediction. Such intuition is also verified in an ablation study.

Then, we use a feed-forward neural network to predict the phrasal NLI label (`Entailment`, `Contradiction`, and `Neutral`). This is given by the standard heuristic matching [43] based on phrase embeddings, followed by a multi-layer perceptron (MLP) and a three-way softmax layer:

$$\begin{aligned} &[P_{\text{phrase}}(\texttt{E}|\text{p}, \text{h}); P_{\text{phrase}}(\texttt{C}|\text{p}, \text{h}); P_{\text{phrase}}(\texttt{N}|\text{p}, \text{h})] \\ &= \text{softmax}(\text{MLP}([\boldsymbol{p}; \boldsymbol{h}; |\boldsymbol{p} - \boldsymbol{h}|; \boldsymbol{p} \circ \boldsymbol{h}])) \end{aligned} \tag{3.3}$$

where $\circ$ is element-wise product and a semicolon refers to column vector concatenation. `E`, `C`, and `N` refer to the `Entailment`, `Contradiction`, and `Neutral` labels, respectively.

Such heuristic matching [43] is proposed to capture the relation between a premise and hypothesis pair, by combining their sentence-level vector representation. Specifically, element-wise difference measures the closeness, while element-wise product

Example 1	
Premise	People are shopping for fruits.
Hypothesis	People are shopping for fruits in the market.
Sentence NLI	[ ] Entailment [ ] Contradiction [✓] Neutral

Example 2	
Premise	People are shopping for fruits in the market.
Hypothesis	People are shopping for fruits.
Sentence NLI	[✓] Entailment [ ] Contradiction [ ] Neutral

Table 3.2: Examples showing the importance of handling unaligned phrases (in highlight).

measures the similarity between them. In our work, we use it in the phrasal level to capture the relation between the aligned phrases.

A multi-layer perceptron (MLP) is a fully-connected feed-forward neural network, which consists of at least three layers: an input layer, a hidden layer and an output layer. In our work, we use the four-layer setting, i.e., an input layer, two hidden layers, and an output layer. And there is a non-linear activation function (namely, the Rectified Linear Unit, or ReLU) between each of the two layers. The input layer takes the combined phrase-level embeddings. Suppose a phrase embedding is k-dimensional, then the combined phrase-level (or heuristically matched) embeddings is $4k$-dimensional (see Equation 3.3). After the MLP, a softmax layer normalizes the logits into probability distribution over `E`, `C`, and `N`.

It should be mentioned that a phrase may be unaligned, but plays an important role in sentence NLI, as shown in Table 3.2. Thus, we would like to predict phrasal NLI labels for unaligned phrases as well, but pair them with a special token ($\mathrm{p}_{\langle \mathrm{EMPTY} \rangle}$ or $\mathrm{h}_{\langle \mathrm{EMPTY} \rangle}$), whose embedding is randomly initialized and learned by back-propagation.

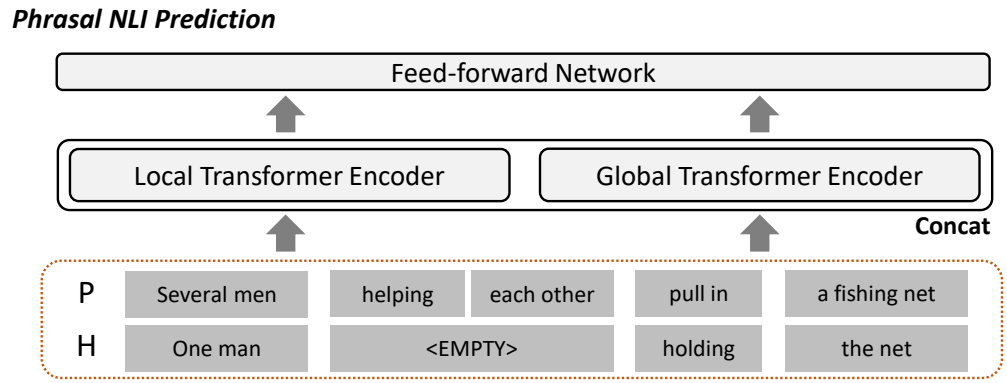

Figure 3.3: Phrase NLI prediction.

3.3 Sentence Label Induction

We observe that the sentence NLI label can be logically induced from phrasal NLI labels. According to the definition of the NLI task, we develop the following induction rules. Figure 3.4 shows the process of the induction from phrasal NLI labels to sentence NLI label.

We utilize Fuzzy Logic [63, 64] to induce phrasal NLI labels to sentence NLI label. Because fuzzy logic models an assertion and performs logic calculation by probability. For example, a quantifier (e.g., "most") and assertion (e.g., "ill") are modeled by a score in $(0, 1)$; the score of a conjunction $s(x_1 \wedge x_2)$ is the product of $s(x_1)$ and $s(x_2)$. In old-school fuzzy logic studies, the mapping from language to the score is usually given by human-defined heuristics [63, 44], and may not be suited to the task of interest. By contrast, we train neural networks in Section 3.2 to predict the probability of phrasal logical relations (`E`, `C` and `N`), and induce the sentence NLI label by fuzzy logic formulas. Thus, our approach takes advantage of both worlds of symbolism and connectionism.

It should be noticed that Fuzzy Logic is not formal Logic, because formal Logic is only applicable to a completely true statement. Whereas Fuzzy Logic can be thought of as a relaxation of Logic, which allows us to perform reasoning with the probability of a true statement.

3.3.1 Entailment Rule

A premise entails a hypothesis, if every paired phrase has the label `Entailment`. Let $\{(\mathrm{p}^{(k)}, \mathrm{h}^{(k)})\}_{k=1}^{K} \bigcup \{(\mathrm{p}^{(k)}, \mathrm{h}^{(k)})\}_{k=K+1}^{K'}$ be all phrase pairs. For $k = 1, \cdots, K$, they are aligned phrases; for $k = K+1, \cdots, K'$, they are unaligned phrases paired with the special token, i.e., $\mathrm{p}^{(k)} = \mathrm{p}_{\langle \mathrm{EMPTY} \rangle}$ or $\mathrm{h}^{(k)} = \mathrm{h}_{\langle \mathrm{EMPTY} \rangle}$. Then, we induce a sentence-level `Entailment` score by

$$S_{\mathrm{sentence}}(\mathtt{E}|\mathrm{P}, \mathrm{H}) = \Big[\prod_{k=1}^{K'} P_{\mathrm{phrase}}(\mathtt{E}|\mathrm{p}^{(k)}, \mathrm{h}^{(k)}) \Big]^{\frac{1}{K'}} \tag{3.4}$$

This works in a fuzzy logic fashion [63, 64], deciding whether the sentence-level label should be `Entailment` considering the average of phrasal predictions. It should be mentioned that, in traditional fuzzy logic, the conjunction is given by the product of probabilities. We find that this gives a too small `Entailment` score compared with `Contradiction` and `Neutral` scores, causing difficulties in end-to-end training. Thus, we take the geometric mean and maintain all the scores in the same magnitude. Here, we use the geometric mean, because it is biased towards low scores, i.e., if there exists one phrase pair with a low `Entailment` score, then the chance of sentence label being `Entailment` is also low. Unaligned pairs should be considered here, because an unaligned phrase may indicate `Entailment`, shown in the second example of Table 3.2. Notice that the resulting value in Equation (3.4) is not normalized with respect to `Contradiction` and `Neutral`; thus, we call it a score (instead of probability), which will be normalized afterwards.

3.3.2 Contradiction Rule

Two sentences are contradictory if there exists (at least) one paired phrase labeled as `Contradiction`. The fuzzy logic version of this induction rule is given by

$$S_{\mathrm{sentence}}(\mathtt{C}|\mathrm{P}, \mathrm{H}) = \max_{k=1,\cdots,K} P_{\mathrm{phrase}}(\mathtt{C}|\mathrm{p}^{(k)}, \mathrm{h}^{(k)}) \tag{3.5}$$

Here, the max operator is used in the induction, because of the contradiction rule is an existential statement, i.e., *there exist(s)* $\cdots$. Also, unaligned phrases are excluded

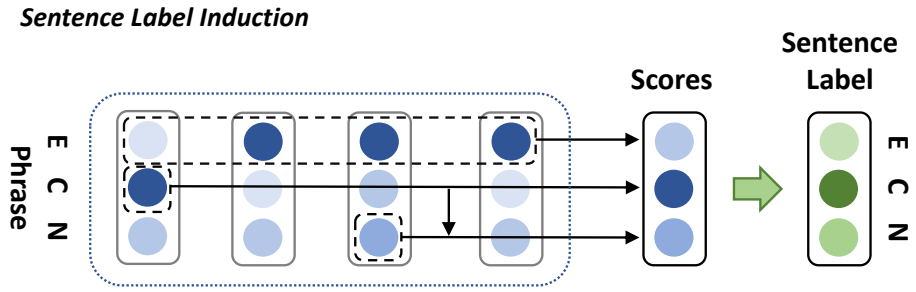

Figure 3.4: Sentence label induction.

in calculating the sentence-level `Contradiction` score, because an unaligned phrase indicates the corresponding information is missing in the other sentence and it cannot be `Contradiction` (recall examples in Table 3.2).

3.3.3 Rule for Neutral

Two sentences are neutral if there exists (at least) one `neutral` phrase pair, but there does not exist any contradictory phrase pair. The fuzzy logic formula is

$$S_{\text{sentence}}(\texttt{N}|\text{P},\text{H}) = \big[\max_{k=1,\cdots,K'} P_{\text{phrase}}(\texttt{N}|\text{p}^{(k)},\text{h}^{(k)})\big] \cdot \big[1 - S_{\text{sentence}}(\texttt{C}|\text{P},\text{H})\big] \tag{3.6}$$

The first factor determines whether there exists a `Neutral` phrase pair (including unaligned phrase, illustrated in the first example in Table 3.2). The second factor evaluates the negation of "at least one contradictory phrase," as suggested in the second clause of the Rule for Neutral.

Finally, we normalize the scores into probabilities by dividing the sum, as all the scores are already positive, given by

$$P_{\text{sentence}}(\texttt{L}|\cdot) = \frac{S_{\text{sentence}}(\texttt{L}|\cdot)}{S_{\text{sentence}}(\texttt{E}|\cdot) + S_{\text{sentence}}(\texttt{C}|\cdot) + S_{\text{sentence}}(\texttt{N}|\cdot)} \tag{3.7}$$

where $\texttt{L} \in \{\texttt{E},\texttt{C},\texttt{N}\}$ is a label.

3.4 Training and Inference

We use cross-entropy loss to train our EPR model by minimizing $-\log P_{\text{sentence}}(\texttt{t}|\cdot)$, where $\texttt{t} \in \{\texttt{E},\texttt{C},\texttt{N}\}$ is the groundtruth sentence-level label.

Our underlying logical reasoning component can be trained end-to-end by back-propagation in a weakly supervised manner, because the fuzzy logic rules are almost everywhere differentiable. While certain points in the max operators in (3.5) and (3.6) may not be differentiable, max operators are common in max-margin learning and the rectified linear unit (ReLU) activation functions, and do not cause trouble in back-propagation.

Once our EPR model is trained, we can obtain both phrase-level and sentence-level labels. This is accomplished by performing argmax on predicted probabilities (3.3) and (3.7), respectively.

3.5 Chapter Summary

In this chapter, we explained the EPR approach that is capable of explaining the model's sentence-level NLI predictions with the phrase-level NLI rationale. EPR is a stack of three components: The first component detects and aligns phrases, the second component makes the phrasal NLI predictions based on the aligned phrases, and the third component induces the phrasal NLI predictions to the final sentence-level prediction that is supervised by the sentence-level NLI labels. Since the model is almost everywhere differentiable, the phrase-level reasoning component can be trained in a weakly supervised manner.

In the next chapter, we will conduct comprehensive experiments for evaluating the performance and explainability of EPR.

Chapter 4

Experiments

In this chapter, we will first introduce our experimental design, including annotating data and developing evaluation metrics. Then, we present experimental results and analysis. We finally provide a case study to show that EPR can perform meaningful reasoning for the NLI task.

4.1 Dataset

We evaluate our EPR approach on the widely used benchmark SNLI dataset [5], which consists of 550K training samples, 10K validation samples, and another 10K test samples. Each data sample consists of two sentences (premise and hypothesis) and a sentence-level groundtruth label.[1]

To evaluate reasoning performance, we need additional human annotation, as no phrasal label is available for NLI reasoning. We performed annotation by three in-lab researchers who are familiar with the NLI task. Our preliminary study shows low agreement when the annotators are unfamiliar with the task; thus it is inappropriate to recruit Mechanic Turkers for annotation.

We select corresponding phrases from both premise and hypothesis, and label them as either `Entailment`, `Contradiction`, or `Neutral`. We may also select a phrase from either a premise or a hypothesis and label it as `Unaligned`. The process can be

[1] A groundtruth label is for a data point, which consists of two sentences. We call it a *sentence-level* label, as opposed to phrasal labels.

repeated until all phrases are labeled for a data sample.

Due to the limit of time and resources, we randomly selected 100 samples for annotation. The amount of annotation follows previous work on textual explanation for SNLI [e-SNLI, 7], and is adequate to show statistical significance. Since our annotation only concerns data samples, it is agnostic to any machine learning model.

We also developed an web annotation interface. For an annotator, the interface consists of several components: a login page, the annotation pages, a navigation panel, and a bookmark page.

We select several pairs of phrases and label them with the NLI relation, by clicking the corresponding buttons (*Entailment*, *Contradiction* or *Neutral*). We may also select a phrase from either a premise or a hypothesis and indicate it is unaligned by clicking the *Unaligned* button. After all phrases are labeled, we need to click *save*, and the information panel will show the result for the phrase-level annotation.

4.2 Development of Evaluation Metrics

For sentence-level NLI prediction, we still use accuracy to evaluate our approach, following previous work [45, 8, 48].

To evaluate phrasal reasoning performance, we need new metrics, because expressive phrasal reasoning is not addressed in any previous work. Specifically, we propose a set of F-scores, which are a balanced measurement of precision and recall between human annotation and model output in terms of `Entailment`, `Contradiction`, `Neutral`, and `Unaligned`.

4.2.1 Precision, Recall, and F-score

F-score is commonly used for evaluating a model's performance on a dataset, which is defined as the the harmonic mean of the model's precision and recall. Considering a binary classification task, computing precision and recall is based on the measurement of four basic elements, which are true positive (TP): when the prediction matches the

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Table 4.1: Confusion matrix.

actual value that are both positive; false positive (FP): when the prediction is positive but the actual value is negative; true negative (TN): when the prediction matches the actual value that are both negative; false negative (FN): when the prediction is negative but the actual value is positive. An confusion matrix is shown in Table 4.1.

Another measurement related to that four basic elements is accuracy, which is defined by

$$\texttt{Accuracy} = \frac{\texttt{TP} + \texttt{TN}}{\texttt{TP} + \texttt{TN} + \texttt{FP} + \texttt{FN}} \tag{4.1}$$

The numerator is the number of correct prediction, and the denominator is the number of total prediction. We choose to use F-score (the combination of precision and recall) is because accuracy is not a good measurement when the dataset is biased to whether positive or negative. For example, a dataset contains 1000 samples, and 900 of them are labeled as positive. So a model can get 90% of accuracy if it simply predicts positive for all samples.

We adopt the measurements of precision and recall to our reasoning metric. Precision is about evaluating the precision of the model within the predicted positives, i.e., how many of the predicted positives are actual positive. It is given by

$$\texttt{Precision} = \frac{\texttt{TP}}{\texttt{TP} + \texttt{FP}} \tag{4.2}$$

Whereas recall calculates the number of the actual positives the model capture within all actual positive samples in the dataset, which is given by

$$\texttt{Recall} = \frac{\texttt{TP}}{\texttt{TP} + \texttt{FN}} \tag{4.3}$$

Example annotation of entailment (in highlight): Premise: A kid in red is playing in a garden. Hypothesis: A child in red is watching TV in the bedroom.								
#	Example Output	$P_{\texttt{E}}^{(P)}$	$P_{E}^{(H)}$	$P_{\texttt{E}}$	$R_{\texttt{E}}^{(P)}$	$R_{E}^{(H)}$	$R_{\texttt{E}}$	$F_{\texttt{E}}$
1	P in a garden H in the bedroom	0	0	0	0	0	0	0
Explanation: Even though *in* occurs in the annotation, the word indexes are different. The reasoning is wrong.								
2	P a kid in red H watching TV	1	0	0	1	0	0	0
Explanation: Mis-matched phrases in hypothesis. The reasoning is wrong.								
3	P a kid | in red H a child | in red	1	1	1	1	1	1	1
Explanation: All word indexes match the annotation. The reasoning is correct.								

Table 4.2: Examples illustrating the proposed metrics, where we consider the `Entailment` category. "|" refers to a phrase segmentation.

A model may have high precision but low recall, or low precision but high recall, when performing on a dataset. F-score takes both precision and recall into consideration so that we have a balanced measurement of the model's performance. The general formula of the weighted F_α-score is given by

$$F_\alpha = (1+\alpha^2) \times \frac{\texttt{Precision} \times \texttt{Recall}}{(\alpha^2 \times \texttt{Precision}) + \texttt{Recall}} \tag{4.4}$$

The factor α is chosen such that recall is considered α times as important as precision. The standard F-score is equivalent to setting α to one, i.e., recall has the same importance as precision. We use standard F-score in our work, which is given by

$$F = \frac{2 \times \texttt{Precision} \times \texttt{Recall}}{\texttt{Precision} + \texttt{Recall}} \tag{4.5}$$

4.2.2 Proposed Reasoning Metric

Consider one sample and the `Entailment` category (Table 4.2). We first count the number of "hits" (true positives) between model output and annotation in terms of word indexes, rather than words. This rules out hitting words in mis-aligned phrases (Example 1, Table 4.2). Then, we calculate precision scores for the premise and hypothesis, denoted by $P_{\texttt{E}}^{(P)}$ and $P_{\texttt{E}}^{(H)}$, respectively. Their geometric mean $P_{\texttt{E}} = (P_{\texttt{E}}^{(P)} P_{\texttt{E}}^{(H)})^{1/2}$ is considered as the precision for `Entailment`. Here, the geometric mean rules out incorrect reasoning that hits either the premise or the hypothesis, but not both (Example 2, Table 4.2). Further, we compute the recall score $R_{\texttt{E}}$ in a similar way, and finally obtain the F-score by $F_{\texttt{E}} = \frac{2P_{\texttt{E}}R_{\texttt{E}}}{P_{\texttt{E}}+R_{\texttt{E}}}$. Likewise, $F_{\texttt{C}}$ and $F_{\texttt{N}}$ are calculated for `Contradiction` and `Neutral`. In addition, we also compute the F-score for unaligned phrases in premise and hypothesis, denoted by $F_{\texttt{UP}}$ and $F_{\texttt{UH}}$, respectively.

When calculating our F-scores for a corpus, we use micro-average, i.e., the precision and recall ratios are calculated in the corpus level. This is more stable, especially considering the varying lengths of sentences. Moreover, we compare model output against three annotators and perform an arithmetic average, further reducing the variance caused by ambiguity.

It should be emphasized that our metrics evaluate phrase detection and alignment in an implicit manner. A poor phrase detector and aligner will result in a low reasoning score (shown in our ablation study), but we do not calculate phrase detection and alignment accuracy explicitly. This helps us cope with the ambiguity of the phrase granularity (Example 3, Table 4.2).

Table 4.3 shows annotation statistics and inter-annotator agreement. As seen, more words are annotated as `Entailment`, whereas fewer are `Contradiction` and `Neutral`. This is understandable, because the two sentences (and thus many phrases) are typically highly related regardless of the sentence-level label, whereas a contra-

	Entailment	Contradiction	Neutral	Unaligned Premise	Hypothesis Hypothesis
Existence (%)	92.33	29.67	29.67	70.00	32.67
Word labels (count)	446.17	100.83	112.33	538	108
Word labels (%)	34.18	7.72	8.61	41.22	8.27
Lower bound (F score)	63.07	18.55	21.41	56.43	24.31
Human performance (F score)	84.71	71.01	55.12	82.46	61.80

Table 4.3: Annotation statistics. The existence measures how often a sentence pair contains a certain label. The lower bound shows the F score of predicting a particular label for all words in the premise and the hypothesis.

dictory/neutral sentence pair only contains one or a few contradictory/neutral phrase pair. Interestingly, the premise contains more unaligned words, whereas the hypothesis contains fewer, which also demonstrates that NLI datasets may be biased [17]. Nevertheless, our EPR model does not make sentence-level predictions by using such bias, because our model is interpretable with explicitly predicted phrasal labels.

We then calculate the "lower-bound" performance by predicting all phrases in the targeted label (`Entailment`, `Contradiction`, `Neutral`, etc.). This is not the performance of a single approach because the F scores are given by respective predicted labels. It in fact demonstrates the difficulty of each label's F score. For example, we achieve 63.07% $F_{\texttt{E}}$ for `Entailment`, but only ˜20% for `Contradition` and `Neutral`.

We also see that humans generally achieve high agreement with each other, which can be thought of "upper-bound" performance. On the contrary, model performance is relatively low (Table 4.4). This shows that our task and metrics are well-defined, yet phrasal reasoning is a challenging task for machine learning models.

Here, the lower and upper bounds are not theoretically guaranteed. They are not strict, but estimate the range of the metric given a reasonable model.

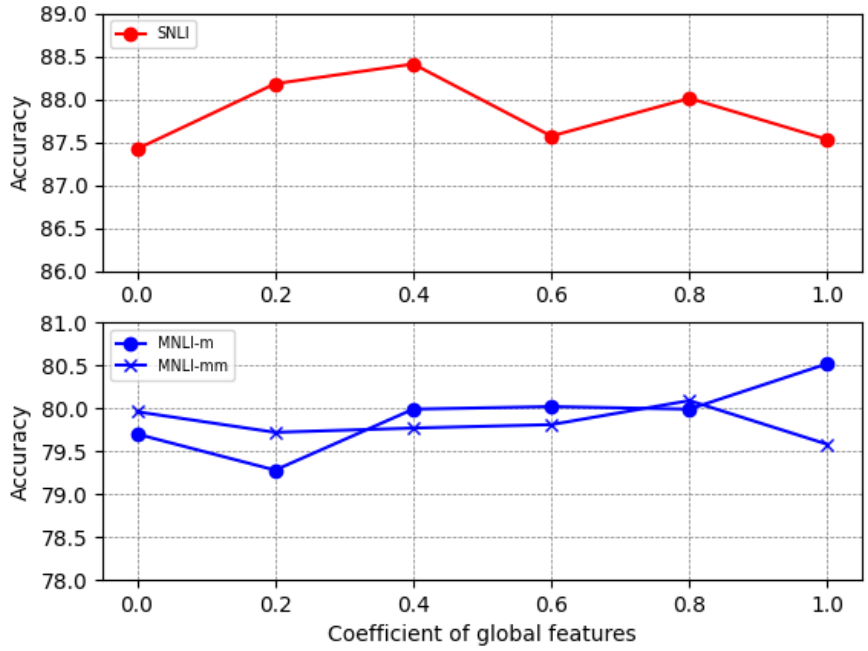

Figure 4.1: Coefficient of global features versus sentence accuracy.

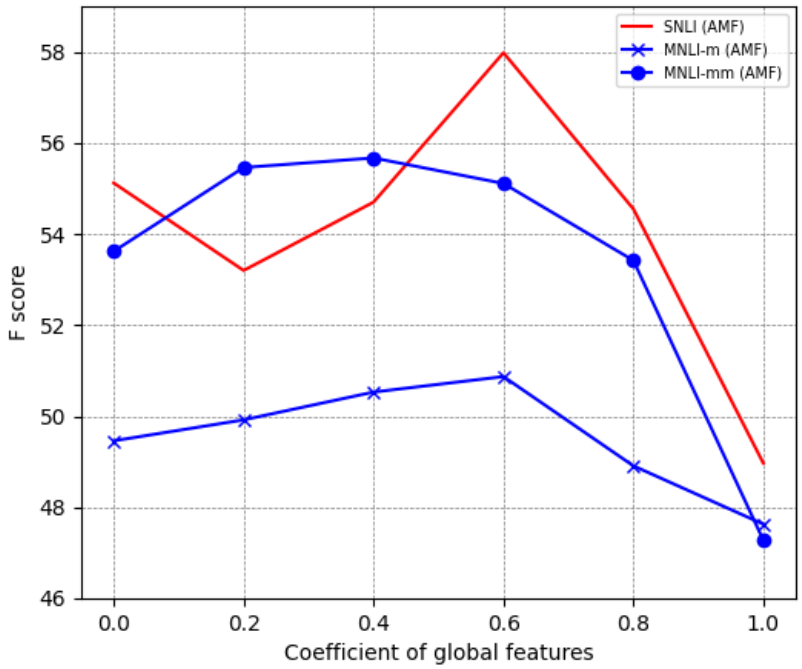

Figure 4.2: Coefficient of global features versus phrasal reasoning performance.

4.3 Training Settings

We chose the pre-trained model `all-mpnet-base-v2`[2] from the Sentence-BERT study [50] and obtained 768-dimensional local and global phrase embeddings. Our MLP had the same dimension as the embeddings, i.e., 768D for the local and global variants, or 1536D for the concatenation variant.

Our alignment model has a hyper-parameter λ (coefficient of global features); we conduct experiment on selecting the best λ from a set of candidates: $\{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, as shown in Figure 4.1 and Figure 4.2. It was set to 0.6, which yields the highest phrasal reasoning performance (we use arithmetic mean F-score for the selection of λ) and decent sentence-level performance on SNLI and Multi-NLI.

During training, the pre-trained language model (LM) is either finetuned or un-finetuned. Fine-tuning yields higher sentence-level accuracy, whereas un-finetuned LM is more efficient for in-depth analyses. We train with a batch size of 256. We used the Adam optimizer [27] with a learning rate of 5e-5, $\beta_1 = 0.9$, $\beta_2 = 0.999$, learning rate warmup over the first 10 percents of the total steps, and linear decay of the learning rate. The model is trained up to 3 epochs, following the common practice [13]. Our main model variants were trained 5 times with different parameter initializations, and we report the mean and standard deviation.

4.4 Main Results

To the best of our knowledge, phrasal reasoning for NLI is not tackled in previous literature. Therefore, we propose plausible extensions to previous studies as our baselines.

We consider the study of Neural Natural Logic [NNL, 16] as the first baseline. It adopts an attention mechanism so that each word in the hypothesis is softly aligned with the words in the premise. Then, each word in the hypothesis is predicted with

[2]https://www.sbert.net/docs/pretrained_models.html

Model	Sent Acc	$F_{\texttt{E}}$	$F_{\texttt{C}}$	$F_{\texttt{N}}$	$F_{\texttt{UP}}$	$F_{\texttt{UH}}$	GM	AM
Human	–	84.71	71.01	55.12	82.46	61.80	70.07	71.02
Non-reasoning								
Mahabadi et al. [40]†	85.1	–	–	–	–	–	–	–
LSTM [56]†	86.1	–	–	–	–	–	–	–
Finetuned Transformer [48]†	**89.9**	–	–	–	–	–	–	–
Baselines								
NNL [16]‡	79.91	62.72	17.49	1.50	66.22	0.00	0.00	29.59
STP	81.44	71.34	36.84	31.09	**76.61**	**51.80**	50.37	53.54
Ours								
EPR (Local, LM unfinetuned)	$76.33_{\pm 0.48}$	$\mathbf{83.11}_{\pm 0.29}$	$38.73_{\pm 0.85}$	$44.63_{\pm 0.88}$	**76.61**	**51.80**	$56.39_{\pm 0.43}$	$58.98_{\pm 0.34}$
EPR (Local, LM finetuned)	$79.36_{\pm 0.13}$	$82.44_{\pm 0.26}$	$\mathbf{44.10}_{\pm 1.32}$	$\mathbf{44.69}_{\pm 3.22}$	**76.61**	**51.80**	$\mathbf{57.77}_{\pm 0.85}$	$\mathbf{59.93}_{\pm 0.67}$
EPR (Concat, LM unfinetuned)	$84.53_{\pm 0.19}$	$73.29_{\pm 0.68}$	$37.95_{\pm 1.16}$	$40.56_{\pm 1.10}$	**76.61**	**51.80**	$53.73_{\pm 0.39}$	$56.04_{\pm 0.33}$
EPR (Concat, LM finetuned)	$\mathbf{87.56}_{\pm 0.15}$	$69.91_{\pm 1.21}$	$39.97_{\pm 2.12}$	$43.31_{\pm 2.78}$	**76.61**	**51.80**	$54.46_{\pm 1.35}$	$56.32_{\pm 1.13}$

Table 4.4: Main results. †Quoted from respective papers. ‡Obtained from the checkpoint sent by the authors. Other results are obtained by our experiments. GM and AM are the geometric and arithmetic means of the F scores.

one of the seven natural logic relations proposed by [37]. We consider the maximum attention score as the alignment, and map their seven natural logic relations to our three-category NLI labels (Table 2.1).

Table 4.4 shows that NNL cannot perform meaningful phrasal reasoning, although our metrics have already excluded explicit evaluation of phrases. The low performance is because their soft attention leads to a large number of mis-alignments, whereas their seven-category logical relations are too fine-grained and cause complications in weakly supervised reasoning. In addition, NNL does not allow unaligned words in the hypothesis, showing that such a model is incapable for NLI reasoning.

By contrast, our EPR model extracts phrases of meaningful semantic units, being an appropriate granularity of logical reasoning. Moreover, we work with three-category NLI labels following the sentence-level NLI task formulation. This actually restricts the model capacity, forcing the model to perform serious phrasal reasoning.

In addition, we include another intuitive BERT-based competing model for comparison. We first apply our own heuristics of phrase detection and alignment (thus,

the model will have the same $F_{\texttt{UP}}$ and $F_{\texttt{UH}}$ scores); then, we directly train the phrasal NLI predictor by sentence-level labels. We call this STP (Sentence label Training Phrases).

As seen, STP provides some meaningful phrasal reasoning results, because the training can smooth out the noise of phrasal labels, which are directly set as the sentence-level labels. But still, its performance is significantly lower than our EPR.

Among our EPR variants, we see that finetuning generally outperforms fixing language models. However, the benefit of finetung is more considarable in sentence-level accuracy than phrasal reasoning.

Moreover, the EPR with local phrase embeddings achieves the highest reasoning performance, and EPR with concatenated features achieve a good balance between sentence-level accuracy and reasoning. Our EPR variants were ran 5 times with different intializations, and standard deviations are also reported in Table 3.3. As seen, our improvement compared with the best baseline is around 8 times of the standard deviation in mean F scores, which is a large margin. Suppose the F scores are Gaussian distributed,[3] the improvement is also statistically significant (p-value <1e-5 comparing our worse variant with the best competing model by one-sided test).

We further compare our EPR with non-reasoning models, which are unable to provide phrasal explainations but may or may not achieve high sentence accuracy. Specifically, Mahabadi et al. [40] apply fuzzy logic to sentence embeddings. They manage to reduce the number of model parameters, but their model is non-interpretable.

4.5 Ablation Study

We conducted an ablation study to verify the effect of every component in our EPR model. We consider three ablated models: (1) Random chunker, which splits the

[3]When the score has a low standard deviation, a Gaussian distribution is a reasonable assumption because its probability of exceeding the range of F scores is extremely low.

Model	Features	Sent Acc	$F_{\texttt{E}}$	$F_{\texttt{C}}$	$F_{\texttt{N}}$	$F_{\texttt{UP}}$	$F_{\texttt{UH}}$	GM	AM
Full model	Local	$76.33_{\pm 0.48}$	$\mathbf{83.11}_{\pm 0.29}$	$\mathbf{38.73}_{\pm 0.85}$	$\mathbf{44.63}_{\pm 0.88}$	**76.61**	**51.80**	$\mathbf{56.39}_{\pm 0.43}$	$\mathbf{58.98}_{\pm 0.34}$
	Global	$84.03_{\pm 0.12}$	$70.84_{\pm 0.60}$	$35.12_{\pm 0.90}$	$36.37_{\pm 1.52}$	**76.61**	**51.80**	$51.41_{\pm 0.62}$	$54.15_{\pm 0.41}$
	Concat	$84.53_{\pm 0.19}$	$73.29_{\pm 0.68}$	$37.95_{\pm 1.16}$	$40.56_{\pm 1.10}$	**76.61**	**51.80**	$53.73_{\pm 0.39}$	$56.04_{\pm 0.33}$
Random chunker	Local	72.44	63.21	22.65	32.04	65.94	36.13	40.53	43.99
	Global	82.81	58.09	30.64	27.49	65.94	36.13	41.05	43.66
	Concat	83.09	58.75	32.41	31.14	65.94	36.13	42.66	44.87
Random alignment	Local	68.52	59.32	21.79	26.20	51.43	16.50	31.02	35.05
	Global	81.99	53.85	35.10	31.39	51.43	16.50	34.71	37.66
	Concat	82.49	57.22	34.83	30.91	51.43	16.50	34.97	38.18
Mean induction	Local	79.61	77.38	37.14	36.13	**76.61**	**51.80**	52.84	55.81
	Global	83.82	55.08	29.92	24.70	**76.61**	**51.80**	43.82	47.62
	Concat	**84.96**	57.12	31.93	31.41	**76.61**	**51.80**	46.92	49.77

Table 4.5: Results of our ablation studies.

sentence randomly based on the number of chunks detected by our system; (2) Random aligner, which randomly aligns phrases but keeps the number of aligned phrases unchanged; and (3) Mean induction, which induces the sentence NLI label by the geometric mean of phrase NLI prediction. In each ablated model, only one component is changed, and other components remain the same as our full EPR model; thus, our comparisons are strictly controlled for drawing scientific conclusions. Due to the large number of settings, each variant was run only once; we do not view this as a concern because the performance gap is large. Also, the underlying language model is un-finetuned, as it yields slightly lower performance but is much more efficient. In addition, we consider local phrase embedding features, global features, and their concatenation in each of the ablated models.

As seen in Table 4.5, the random chunker and aligner yield poor phrasal reasoning performance, showing that working with meaningful semantic units and having their relationship are important to logical reasoning. This also verifies that our metrics are able to evaluate phrase detection and alignment in an implicit manner.

Interestingly, local features yield higher reasoning performance, but global and concatenated features yield higher sentence accuracy. This is because global fea-

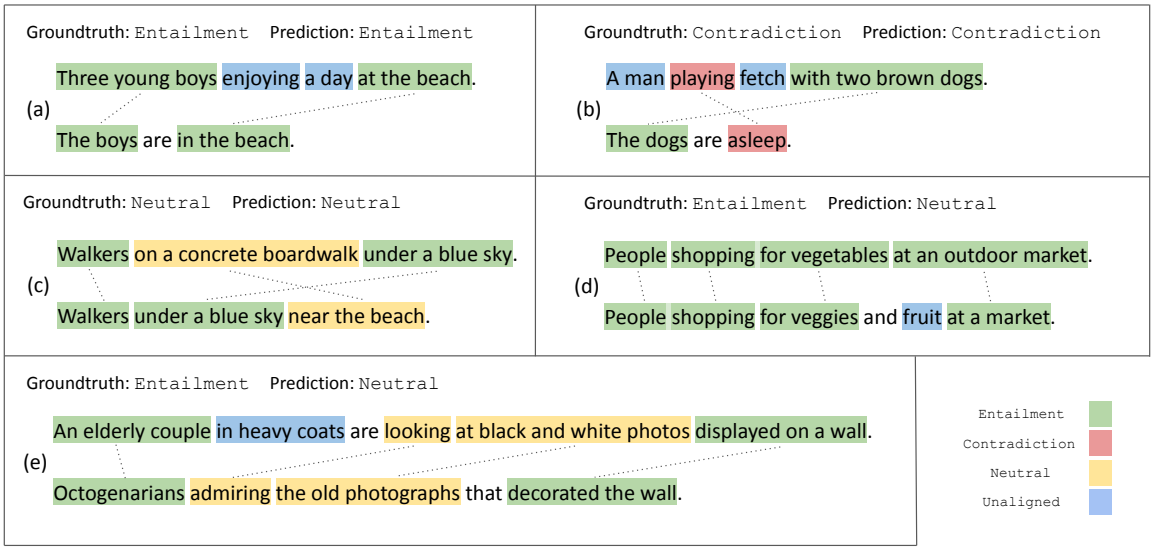

Figure 4.3: Examples of explainable phrasal reasoning predicted by our EPR model. Words in one color block are a detected phrase; a dotted line shows the alignment of two phrases; and the color represents the predicted phrasal NLI label. In Example (d) and (e), EPR's prediction suggests provided label in SNLI is incorrect.

tures provide aggregated information of the entire sentence, but also allow bypassing meaningful reasoning. In the variant of the mean induction, for example, the phrasal predictor can simply learn to predict the sentence-level label with global sentence information; then, the mean induction is an ensemble of multiple predictors. In this way, it achieves the highest sentence accuracy (0.43 points higher than our full model), but is 6 points lower in reasoning performance.

This reminds us of the debate between old schools of AI [52]. Recent deep learning models take the connectionists' view, and generally outperform symbolists' approaches in terms of the ultimate prediction, but they lack expressible explanations. Combining neural and symbolic methods becomes a hot direction in recent AI research [31, 14, 61]. In general, our EPR model with global features achieves high performance in both reasoning and ultimate prediction for the NLI task.

4.6 Case Study

We present case studies in Figure 4.3. We see that our EPR indeed performs impressive reasoning for the NLI task, which is learned in a weakly supervised manner. In Example (a), the two sentences are predicted `Entailment` because *three young boys* entails *the boys* and *at the beach* entails *in the beach*, whereas unaligned phrases *enjoying* and *a day* are allowed in the premise for `Entailment`. In Example (b), *playing* contradicts *asleep*, and the two sentences are also predicted `Contradiction`. Likewise, Example (c) is predicted `Neutral` because the aligned phrases *on a concrete boardwalk* and *near the beach* are neutral.

In our study, we also find several interesting examples where EPR's reasoning provides clues suggesting that the target labels may be incorrect in the SNLI dataset. In Example (d), the target label is `Entailment`. However, our EPR model determines that the two sentences should be `Neutral` because *fruit* is unaligned in the hypothesis. Indeed, we believe our model's reasoning and prediction make more sense than the provided target label, because people shopping for vegetables may or may not shop for fruit. In Example (e), our model predicts `Neutral` for *looking* and *admiring*, as well as for *at black and white photos* and *the old photographs*. Thus, the two sentences are predicted `Neurtral`, as opposed to the provided label `Contradiction`. We believe our model's reasoning and prediction are correct, because people looking at something may or may not admire it; a black-and-white photo may or may not be an old photo either (as it could be a black-and-white artistic photo).

4.7 Additional Experiment on MNLI

We provide additional results on the Multi-NLI (MNLI) dataset [57], which is much noisier compared with SNLI. We nevertheless conducted additional experiments on MNLI to provide further evidence of our EPR approach. MNLI contains two sections of test sets: the matched and mismatched sections. For the matched section, we

aim to show statistically significance following Section 4.4. Whereas the results on mismatched section shows the robustness of our method.

4.7.1 Dataset and Annotation

MNLI consists of 393K training samples, 20K validation samples and another 20K test samples (10K for matched and 10K for mismatched). It has the same format as SNLI dataset, but samples come from multiple domains and are more diverse. We follow Section 4.1 and use the same protocol to create the phrasal reasoning annotation for MNLI dataset based on 100 randomly selected samples. However, we found that MNLI is much noisier than SNLI; particularly, the sentences labeled as `Neutral` in MNLI share few relevant phrases. For example, the two sentences do not have much in common in the sample "*Premise: If you still want to join, it might be worked.*" and "*Hypothesis: Your membership is the only way that this could work*". Moreover, inter-human agreement is low in terms of the `Neutral` category. Therefore, we believe the corpus quality is low for `Neutral`. To ensure meaningful evaluation, we ignored the evaluation of `Neutral` in this experiment, although our reasoning approach is not changed. The remaining 60 samples containing `Entailment` and `Contradiction` serve as the MNLI phrasal reasoning corpus. Following the annotation analysis for SNLI in Table 4.3, we show annotation statistics and inter-annotator agreement for the matched and mismatched dataset section in Table 4.6 and Table 4.7, respectively.

4.7.2 Results on MNLI

We consider the EPR variant with concatenated local and global features, since the SNLI experiment shows it achieves a good balance between sentence-level accuracy and reasoning. Our models were run 5 times with different intializations, and inference on the MNLI matched test set to report mean and standard deviation. While the model was run only one time on the MNLI mismatched test set to show the robustness of our approach.

	Entailment	Contradiction	Unaligned Premise	Unaligned Hypothesis
Existence (%)	42.33	24	29.67	12.33
Word labels (count)	202.67	86.83	133.67	28.33
Word labels (%)	44.89	19.23	29.61	6.28
Lower bound (F score)	64.70	34.98	42.91	12.36
Human performance (F score)	85.15	73.44	73.18	46.31

Table 4.6: Annotation statistics for MNLI-matched. The existence measures how often a sentence pair contains a certain label. The lower bound shows the F score of predicting a particular label for all words in the premise and the hypothesis.

	Entailment	Contradiction	Unaligned Premise	Unaligned Hypothesis
Existence (%)	46.67	24.33	29.67	23
Word labels (count)	229	64	117.67	67
Word labels (%)	48.04	13.22	24.69	14.06
Lower bound (F score)	70.99	26.04	38.98	25.97
Human performance (F score)	82.54	69.74	68.94	54.29

Table 4.7: Annotation statistics for MNLI-mismatched.

Model	Sent Acc	$F_{\texttt{E}}$	$F_{\texttt{C}}$	$F_{\texttt{UP}}$	$F_{\texttt{UH}}$	GM	AM
Human	–	85.15	73.44	73.18	46.31	67.85	69.52
Non-reasoning methods							
Mahabadi et al. [40]†	73.8	–	–	–	–	–	–
Multi-task BiLSTM + Attn [55]†	72.2	–	–	–	–	–	–
Finetuned Transformer [48]†	**82.1**	–	–	–	–	–	–
Reasoning methods							
NNL [16]‡	61.28	50.33	32.00	49.78	0.00	0.00	33.03
STP	64.46	58.01	34.79	**64.32**	**37.57**	46.99	48.67
EPR (Concat, LM finetuned)	**79.65**$_{\pm 0.19}$	**61.76**$_{\pm 0.32}$	**52.09**$_{\pm 0.41}$	**64.32**	**37.57**	**52.80**$_{\pm 0.07}$	**53.93**$_{\pm 0.07}$

Table 4.8: Results on MNLI. †Quoted from respective papers. ‡Our replication.

Model	Sent Acc	$F_{\texttt{E}}$	$F_{\texttt{C}}$	$F_{\texttt{UP}}$	$F_{\texttt{UH}}$	GM	AM
Human	–	82.54	69.74	68.94	54.29		
Non-reasoning methods							
Mahabadi et al. [40]†	73.7	–	–	–	–	–	–
Multi-task BiLSTM + Attn [55]†	72.1	–	–	–	–	–	–
Finetuned Transformer [48]†	**81.4**	–	–	–	–	–	–
Reasoning methods							
NNL [16]‡	61.38	63.34	16.37	45.93	0.00	0.00	31.41
STP	65.07	**70.04**	34.04	**64.65**	**52.23**	53.27	55.24
EPR (Concat, LM finetuned)	**79.81**	67.03	**39.12**	**64.65**	**52.23**	**54.55**	**55.76**

Table 4.9: Results on MNLI (mis-matched). †Quoted from respective papers. ‡Our replication.

As seen in Table 4.8, our EPR approach is again worse than humans, but largely improves the reasoning performance compared with NNL and STP baselines. Its sentence-level prediction is also comparable to (although slightly lower than) fine-tuning Transformers. Again, the improvement is also statistically significant (p-value <1e-5 comparing our worse variant with the best competing model by one-sided test), which is highly consistent with SNLI experiments.

Results on MNLI-mismatched in Table 4.9 further verify that our approach is

robust. Our EPR model has the best overall phrase-level reasoning ability, as well as the best sentence-level accuracy among all the baselines. However, we found that the `Entailment` F-score of our EPR model is slightly lower than the best baseline STP. This is because the corpus is biased to the annotated phrases in the `Entailment` category (Table 4.7 shows that entailed words count almost half of the words in corpus). The lower bound F-score is very high, and even higher than the performance of the baselines and our EPR model. This indicates that simply making use of the corpus' bias toward `Entailment` can achieve the highest reasoning performance in `Entailment` among all reasoning methods. However, we believe the EPR model does not take such bias because of the nature of fuzzy logic formulas. Reasoning in `Contradiction` category is more difficult because the `Contradiction` words only account for 13.22% of the words in corpus, but the EPR model has the highest `Contradiction` F-score, which shows the robustness of our method.

4.8 Chapter Summary

In this chapter, we presented an evaluation metric based on F-score and the dataset annotation process for the NLI reasoning. We annotated the reasoning datasets for two widely used NLI benchmark datasets: SNLI and MNLI. The experiments show that, compared with the baselines, EPR boosts the reasoning performance on both NLI datasets. EPR's sentence-level performance is comparable to a fine-tuned Transformer model, but the latter is not capable of reasoning. We also conducted an ablation study to verify the effectiveness of each EPR component. The results of different model variants using local and global embeddings provide insights into the neuro-symbolic method, based on which we observe that introducing symbolic knowledge to a neural network may slightly hurt the model's generalization ability, but helps explainability and interpretability.

Chapter 5

Conclusion

5.1 Thesis Summary

Natural Language Inference (NLI) is a fundamental task in natural language processing. There are a large number of studies that propose new methods for improving the performance of the NLI task. However, such improvement is mostly because of the success of the large pre-train language models. Although there are some studies addressing the reasoning for NLI, they lack an explicit evaluation metrics for evaluating the reasoning ability for their proposed systems.

The thesis proposes an explainable phrasal reasoning (EPR) model for natural language inference (NLI). Our EPR first detects and aligns meaningful semantic units (roughly speaking, phrases) as the granularity of reasoning. Then, EPR predicts phrasal NLI labels and induces them to the sentence level by fuzzy logic; our reasoning component can be trained in a weakly supervised manner, as it is almost everywhere differentiable.

To evaluate our approach, we proposed an experimental design, including data annotation, evaluation metrics, and plausible baselines. Results show that phrasal reasoning for NLI is a meaningfully defined task, as humans can achieve high agreements. Our EPR achieves decent sentence-level accuracy but much higher reasoning performance than all competing models.

Case studies provide qualitative evidence that EPR indeed performs meaningful

reasoning for the NLI task; based on EPR's predicted explanation, we are able to detect incorrect target labels in the original SNLI dataset.

5.2 Limitations and Future Work

Our method performs phrase detection and alignment by heuristics. These rules and heuristics work well empirically in our experiments, although they can be further improved, for example, by considering syntactic features. Unsupervised methods may also be applied for phrase detection [11]. However, our main focus is neural fuzzy logic for weakly supervised reasoning. This largely differs from previous work based on manually designed lexicons and rules [21, 9].

Our long-term goal is to develop a weakly supervised, end-to-end trained neuro-symbolic system that can extract semantic units and perform reasoning for a given downstream NLP task.

Currently, we manually define the fuzzy logic induction formulas, based on our knowledge of phrase-level NLI reasoning. It is also possible to let the system extract or learn the fuzzy logic formulas automatically from a given NLP task. Therefore, our ultimate desired reasoning system is capable of fully automatic self-reasoning. And this thesis is an important milestone toward the long-term goal.

Bibliography

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. "Layer normalization". In: *arXiv preprint arXiv:1607.06450* (2016).

[2] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *International Conference on Learning Representations*. 2015.

[3] Islam Beltagy, Stephen Roller, Pengxiang Cheng, Katrin Erk, and Raymond J Mooney. "Representing meaning with a combination of logical and distributional models". In: *Computational Linguistics* 42.4 (2016), pp. 763–808.

[4] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. "A neural probabilistic language model". In: *Advances in Neural Information Processing systems* 13 (2000).

[5] Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. "A large annotated corpus for learning natural language inference". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015, pp. 632–642. URL: https://aclanthology.org/D15-1075.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. "Language models are few-shot learners". In: *Advances in Neural Information Processing systems* 33 (2020), pp. 1877–1901.

[7] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. "e-SNLI: Natural language inference with natural language explanations". In: *Advances in Neural Information Processing Systems*. 2018, pp. 9539–9549.

[8] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. "Enhanced LSTM for natural language inference". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2017, pp. 1657–1668. URL: https://aclanthology.org/P17-1152/.

[9] Zeming Chen, Qiyue Gao, and Lawrence S Moss. "NeuralLog: Natural language inference with joint neural and logical Reasoning". In: *arXiv preprint arXiv:2105.14167* (2021). URL: https://arxiv.org/abs/2105.14167.

[10] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning phrase representations using RNN encoder–decoder for statistical machine Translation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2014, pp. 1724–1734. URL: https://aclanthology.org/D14-1179.

[11] Anup Anand Deshmukh, Qianqiu Zhang, Ming Li, Jimmy Lin, and Lili Mou. "Unsupervised chunking as syntactic structure induction with a knowledge-transfer approach". In: *Findings of the Association for Computational Linguistics: EMNLP*. 2021, pp. 3626–3634.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "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, Volume 1 (Long and Short Papers)*. 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423.

[13] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. "Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping". In: *arXiv preprint arXiv:2002.06305* (2020). URL: https://arxiv.org/abs/2002.06305.

[14] Honghua Dong, Jiayuan Mao, Tian Lin, Chong Wang, Lihong Li, and Denny Zhou. "Neural logic machines". In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=B1xY-hRctX.

[15] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. "Explainable artificial intelligence: A survey". In: *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics*. IEEE. 2018, pp. 0210–0215.

[16] Yufei Feng, Quan Liu, Michael Greenspan, Xiaodan Zhu, et al. "Exploring end-to-end differentiable natural logic modeling". In: *Proceedings of the 28th International Conference on Computational Linguistics*. 2020, pp. 1172–1185. URL: https://aclanthology.org/2020.coling-main.101.

[17] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. "Annotation artifacts in natural language inference data". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 2018, pp. 107–112. URL: https://aclanthology.org/N18-2017/.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778.

[19] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[20] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. *spaCy: Industrial-strength natural language processing in Python*. 2020. URL: https://doi.org/10.5281/zenodo.1212303.

[21] Hai Hu, Qi Chen, Kyle Richardson, Atreyee Mukherjee, Lawrence S Moss, and Sandra Kübler. "MonaLog: A lightweight system for natural language inference based on monotonicity". In: *Proceedings of the Society for Computation in Linguistics*. 2020, pp. 284–293. URL: https://aclanthology.org/2020.scil-1.40/.

[22] Minghao Hu, Yuxing Peng, Zhen Huang, Xipeng Qiu, Furu Wei, and Ming Zhou. "Reinforced mnemonic reader for machine reading Comprehension". In: *arXiv preprint arXiv:1705.02798* (2017).

[23] Eric Jang, Shixiang Gu, and Ben Poole. "Categorical reparameterization with Gumbel-Softmax". In: *International Conference on Learning Representations*. 2017. URL: https://openreview.net/forum?id=rkE3y85ee.

[24] Zhongtao Jiang, Yuanzhe Zhang, Zhao Yang, Jun Zhao, and Kang Liu. "Alignment rationale for natural language inference". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021, pp. 5372–5387. URL: https://aclanthology.org/2021.acl-long.417/.

[25] Kun Jing and Jungang Xu. "A survey on neural network language models". In: *arXiv preprint arXiv:1906.03591* (2019).

[26] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. "Dense passage retrieval for open-domain question answering". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 2020, pp. 6769–6781. URL: https://aclanthology.org/2020.emnlp-main.550/.

[27] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014). URL: https://arxiv.org/abs/1412.6980.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in Neural Information Processing systems* 25 (2012).

[29] Tao Lei, Regina Barzilay, and Tommi Jaakkola. "Rationalizing neural predictions". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2016, pp. 107–117. URL: https://aclanthology.org/D16-1011/.

[30] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension". In: *arXiv preprint arXiv:1910.13461* (2019).

[31] Chen Liang, Jonathan Berant, Quoc Le, Kenneth Forbus, and Ni Lao. "Neural symbolic machines: learning semantic parsers on Freebase with weak supervision". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2017, pp. 23–33. URL: https://aclanthology.org/P17-1003/.

[32] Xianggen Liu, Lili Mou, Haotian Cui, Zhengdong Lu, and Sen Song. "Finding decision jumps in text classification". In: *Neurocomputing* 371 (2020), pp. 177–187. URL: https://doi.org/10.1016/j.neucom.2019.08.082.

[33] Xianggen Liu, Lili Mou, Haotian Cui, Zhengdong Lu, and Sen Song. "Jumper: Learning when to make classification decisions in reading". In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. 2018, pp. 4237–4243. URL: https://dl.acm.org/doi/10.5555/3304222.3304359.

[34] Yang Liu and Mirella Lapata. "Text summarization with pretrained encoders". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 2019, pp. 3730–3740. URL: https://aclanthology.org/D19-1387/.

[35] Zhengdong Lu, Xianggen Liu, Haotian Cui, Yukun Yan, and Daqi Zheng. "Object-oriented neural programming (OONP) for Document Understanding". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 2018, pp. 2717–2726. URL: https://aclanthology.org/P18-1253.

[36] Bill MacCartney, Michel Galley, and Christopher D Manning. "A phrase-based alignment model for natural language inference". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. 2008, pp. 802–811.

[37] Bill MacCartney and Christopher D Manning. "An extended model of natural logic". In: *Proceedings of the Eighth International Conference on Computational Semantics*. 2009, pp. 140–156. URL: https://aclanthology.org/W09-3714.

[38] Bill MacCartney and Christopher D Manning. "Natural logic for textual inference". In: *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. 2007, pp. 193–200. URL: https://aclanthology.org/W07-1431/.

[39] Bill MacCartney and Christopher D. Manning. "Modeling semantic containment and exclusion in natural language inference". In: *Proceedings of the 22nd International Conference on Computational Linguistics*. 2008, pp. 521–528. URL: https://aclanthology.org/C08-1066.

[40] Rabeeh Karimi Mahabadi, Florian Mai, and James Henderson. "Learning entailment-based sentence embeddings from natural language inference". In: *Online Manuscript* (2019). URL: https://openreview.net/forum?id=BkxackSKvH.

[41] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[42] Lili Mou, Zhengdong Lu, Hang Li, and Zhi Jin. "Coupling distributed and symbolic execution for natural language queries". In: *International Conference on Machine Learning*. 2017, pp. 2518–2526.

[43] Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. "Natural language inference by tree-Based convolution and heuristic matching". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2016, pp. 130–136. URL: https://aclanthology.org/P16-2022.

[44] Ken Nozaki, Hisao Ishibuchi, and Hideo Tanaka. "A simple but powerful heuristic method for generating fuzzy rules from numerical data". In: *Fuzzy Sets and Systems* 86.3 (1997), pp. 251–270. URL: https://www.sciencedirect.com/science/article/abs/pii/0165011495004130.

[45] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. "A decomposable attention model for natural language inference". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2016, pp. 2249–2255. URL: https://aclanthology.org/D16-1244/.

[46] Slav Petrov, Dipanjan Das, and Ryan McDonald. "A universal part-of-speech tagset". In: *arXiv preprint arXiv:1104.2086* (2011).

[47] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. "Hypothesis only baselines in natural language inference". In: *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, 2018, pp. 180–191. URL: https://aclanthology.org/S18-2023.

[48] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. "Improving language understanding by generative pre-training". In: *OpenAI Blog* (2018).

[49] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. "Squad: 100,000+ questions for machine comprehension of text". In: *arXiv preprint arXiv:1606.05250* (2016).

[50] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence embeddings using siamese BERT-networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 2019. URL: http://arxiv.org/abs/1908.10084.

[51] Cynthia Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215. URL: https://www.nature.com/articles/s42256-019-0048-x.

[52] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach (Third Edition)*. Pearson Education Ltd., 2016.

[53] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 618–626.

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In: *Advances in Neural Information Processing Systems*. 2017, pp. 5998–6008.

[55] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. "GLUE: A multi-task benchmark and analysis platform for natural language understanding". In: *arXiv preprint arXiv:1804.07461* (2018).

[56] Shuohang Wang and Jing Jiang. "Learning natural language inference with LSTM". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2016, pp. 1442–1451. URL: https://aclanthology.org/N16-1170/.

[57] Adina Williams, Nikita Nangia, and Samuel Bowman. "A broad-coverage challenge corpus for sentence understanding through inference". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2018.

[58] Wenhan Xiong, Thien Hoang, and William Yang Wang. "DeepPath: A reinforcement learning method for knowledge graph reasoning". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017, pp. 564–573. URL: https://aclanthology.org/D17-1060/.

[59] Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. "Can neural networks understand monotonicity reasoning?" In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 2019, pp. 31–40. URL: https://aclanthology.org/W19-4804.

[60] Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. "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 2019)*. 2019, pp. 250–255. URL: https://aclanthology.org/S19-1027.

[61] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. "Neural-symbolic VQA: Disentangling reasoning from vision and language understanding". In: *Advances in Neural Information Processing Systems*. 2018. URL: https://dl.acm.org/doi/10.5555/3326943.3327039.

[62] Deunsol Yoon, Dongbok Lee, and SangKeun Lee. "Dynamic self-attention: Computing attention over words dynamically for sentence embedding". In: *arXiv preprint arXiv:1808.07383* (2018). URL: https://arxiv.org/abs/1808.07383.

[63] Lotfi A Zadeh. "Fuzzy logic". In: *Computer* 21.4 (1988), pp. 83–93.

[64] Lotfi A Zadeh. "Fuzzy sets". In: *Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems*. World Scientific, 1996, pp. 394–432. URL: https://www.worldscientific.com/doi/abs/10.1142/9789814261302_0021.

[65] Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. "Semantics-aware BERT for language understanding". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020, pp. 9628–9635. URL: https://doi.org/10.1609/aaai.v34i05.6510.