ERNIE-NLI: Analyzing the Impact of Domain-Specific External Knowledge on Enhanced Representations for NLI

Lisa Bauer[1], **Lingjia Deng**[2], **Mohit Bansal**[1]
[1]UNC Chapel Hill [2]Bloomberg
{lbauer6, mbansal}@cs.unc.edu
{ldeng43}@bloomberg.net

Abstract

We examine the effect of domain-specific external knowledge variations on deep large scale language model performance. Recent work in enhancing BERT with external knowledge has been very popular, resulting in models such as ERNIE (Zhang et al., 2019a). Using the ERNIE architecture, we provide a detailed analysis on the types of knowledge that result in a performance increase on the Natural Language Inference (NLI) task, specifically on the Multi-Genre Natural Language Inference Corpus (MNLI). While ERNIE uses general TransE embeddings, we instead train domain-specific knowledge embeddings and insert this knowledge via an information fusion layer in the ERNIE architecture, allowing us to directly control and analyze knowledge input. Using several different knowledge training objectives, sources of knowledge, and knowledge ablations, we find a strong correlation between knowledge and classification labels within the same polarity, illustrating that knowledge polarity is an important feature in predicting entailment. We also perform classification change analysis across different knowledge variations to illustrate the importance of selecting appropriate knowledge input regarding content and polarity, and show representative examples of these changes.

1 Introduction

Recently, the selection and integration of external knowledge into large-scale language models has shown impressive improvements in several Natural Language Understanding (NLU) tasks (Zhang et al., 2019a). Understanding the relation between external knowledge and model performance is fundamental to understanding how best to select and integrate knowledge into NLU tasks. We focus specifically on Natural Language Inference (NLI), which requires understanding sentence semantics with respect to both the content and polarity. NLI is motivated by recognizing textual entailment, or understanding whether a hypothesis entails, contradicts, or is neutral with respect to a premise. For example, given the premise: "Some boys are playing soccer", the hypothesis "Young men are playing a sport" is an entailment whereas the hypothesis "Old men are playing a sport" is a contradiction. Language modeling is a very common and important approach when considering the NLI task.

The NLI state-of-the-art utilizes different language modeling techniques to learn the relations between the hypothesis and the premise. Yoon et al. (2018) used Dynamic Self-Attention (DSA) to learn sentence embeddings, Liu et al. (2019) proposed multi-task deep neural network (MT-DNN) for learning language representations in multiple NLU tasks, and Zhang et al. (2019b) combined semantic role labeling and BERT (Devlin et al., 2019) to explicitly absorb contextual semantics over a BERT framework. However, these approaches limit the source of information available for representing both the premise and hypothesis. Consider the following premise and hypothesis:

People cut their expenses for the Golden years.

People decrease their expenses for retirement.

It is challenging to know that "Golden years" entails "retirement" if we rely only on the context within the two sentences. To illustrate how common this problem is, we conduct a manual analysis of BERT classification errors on the NLI task (specifically on the MNLI corpus (Williams et al., 2018), more details in Section 6), and find that at least 50% of misclassifications require external knowledge, specifically requiring domain-specific knowledge, world knowledge, jargon-based paraphrases, or commonsense knowledge to resolve the entailment. In the above example, a model that learns the relation between "Golden years" and "retirement" from external knowledge can be used to enhance NLI inference.

On the basis of this idea, Chen et al. (2018) and Zhang et al. (2019a) used external knowledge from

Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, pages 58–69
Online, June 10, 2021. ©2021 Association for Computational Linguistics

WordNet and TransE (Bordes et al., 2013) and applied it to NLI models. In their work, pre-trained representations of external knowledge from knowledge bases (e.g., TransE) were directly applied; they did not tailor knowledge content or structure specifically to the NLI task and did not improve NLI performance (Zhang et al., 2019a). This finding motivates our investigation on how external knowledge can be efficiently used to improve NLI models. The intention of our work is not to propose a new model that outperforms the state-of-the-art, but instead to focus on building a framework for investigating how different types and representations of external knowledge impact an NLI model's decisions.

Consider our previous examples. We want to represent that the relation between "young men" and "boys" is positive for entailment, and that the relation between "old men" and "boys" is negative for entailment. Similarly, we want to represent that the relation between "Golden years" and "retirement" is positive for entailment. The interplay of external knowledge and entailment gives insight into the power of selecting relevant knowledge with respect to both content and polarity of the knowledge. Here, content indicates the semantic meaning of external knowledge and polarity indicates whether the knowledge relation is positive or negative for entailment. The representation of external knowledge is required to be correct in both aspects for the NLI task. The models learns (1) content via our knowledge extraction phase, by extracting concept edges from knowledge graphs, and (2) polarity via our knowledge training phase, by learning the polarity of the relationships between concepts. We define concepts as words or phrases througout this paper. In this work, we aim to show what type of external knowledge is useful for certain classes of NLI. We examine how different types of knowledge impact neural language model decisions with respect to content and polarity.

To this end, we propose ERNIE-NLI, an NLI model that integrates external knowledge to enhance and probe NLI inference decisions. First, we adapt knowledge content in various sources to our setup: external knowledge relations are mapped to NLI knowledge relations (Section 4.2). In this step, we not only represent external knowledge from different sources in a unified way, but also convert external knowledge content to the NLI task. Second, the polarity is learned (Section 4.3): NLI knowledge embeddings are learned to predict whether they are positive or negative for entailment. In this step, we extend BERT with a knowledge embedding layer and a classification layer. Third, the content and polarity are applied to NLI classification (Section 4.4). All three phases listed above are depicted in Fig. 1. ERNIE-NLI is developed on the basis of ERNIE (Zhang et al., 2019a), which did not improve performance on the NLI task, although it was infused with TransE embeddings. Results show that our model ERNIE-NLI enhanced with adapted knowledge achieves better performance than ERNIE for specific classes depending on knowledge input.

We perform an in-depth analysis to examine how different types of knowledge impact NLI model's decisions with respect to content and polarity. We conduct a series of experiments to investigate why and how the adapted knowledge enhances NLI predictions. From the experiments, we find that:

- Integrating knowledge improves performance for NLI classes that correspond to integrated knowledge with regards to the polarity (e.g., positive knowledge improves entailment classification, etc.).
- Increased amount of knowledge during training improves performance for NLI labels that correspond to increased knowledge with regards to the polarity.
- Presence of knowledge at inference improves performance for NLI labels that correspond to present knowledge with regards to polarity (e.g., a correct entailment prediction with the presence of positive knowledge is observed to occur more often than with the presence of negative knowledge, etc.).
- ERNIE-NLI performance is robust to new knowledge content.

In summary, the proposed NLI model enhanced with adapted external knowledge from various sources achieves better performance for respective classes, allows us to analyze the impact of knowledge type, and is robust when the knowledge at inference time has shifted. We examine this performance with detailed analysis throughout the paper. Overall our contributions are as follows:

- We propose a knowledge analysis framework, ERNIE-NLI, that allow us to directly control and analyze adapted knowledge input, to investigate the characteristics of knowledge that result in a performance increase on the NLI task.

- We present findings that show strong correlations between knowledge polarity and downstream performance, illustrating the knowledge features that are important for increased performance.
- We perform extensive analysis and experimentation to support our findings (e.g., classification change analysis, adding knowledge incrementally, adding unseen knowledge, etc).

2 Related Work

2.1 Natural Language Inference

Early work in Natural Language Inference, also known as Textual Entailment (Dagan et al., 2005), exploited different features including logical rules (Bos and Markert, 2005), dependency parsers (Iftene and Balahur, 2007), and semantics (MacCartney and Manning, 2009), etc. With the development of large human annotated corpus such as the Stanford Natural Language Inference Corpus (Bowman et al., 2015) and the Multi-Genre NLI Corpus (Williams et al., 2018), most recent work has explored various neural models.

Different encoders have been studied to represent sentences, including LSTM (Bowman et al., 2016), tree-based CNN (Mou et al., 2015), TreeLSTM (Choi et al., 2018), etc. Previous work has explored using dynamic self-attention (Yoon et al., 2018), distance-based self-attention (Im and Cho, 2017) and reinforced self-attention (Shen et al., 2018) to enhance sentence encoders. Ensemble methods that combine multiple models have also shown improvements (Wang et al., 2017; Peters et al., 2018; Kim et al., 2019). Sun et al. (2019) improved masked language modeling with knowledge masking strategies, via entity-level and phrase-level masking, which showed improvement on NLI. Sun et al. (2020) then expanded this work to continual pre-training, which incrementally learns pre-training tasks through constant multi-task learning. Peters et al. (2019) investigated embedding knowledge bases into large-scale models in a multitask setup, seeing improvements on relationship extraction, entity typing, and word sense disambiguation.

Using external knowledge to enhance NLI models specifically, Chen et al. (2018) obtained the semantic relations between words from WordNet and calculated the relation embeddings using pre-trained TransE embeddings. Additionally, previous work has explored injecting lexical knowledge into pre-trained models for MNLI (Williams et al., 2018), among other tasks (Lauscher et al., 2020; Levine et al., 2020). Zhang et al. (2019a) adopted a knowledgeable encoder to inject the knowledge information into language representation. However, in contrast to our work, their external knowledge was not trained specifically for the NLI task.

2.2 Knowledge Embeddings

Using knowledge embeddings that represent the relations between entities has been useful in various downstream NLP tasks. Bordes et al. (2013) proposed TransE, a method which modeled relationships by interpreting them as translations operating on the low-dimensional embeddings of the entities. To address the issue of complex relation embeddings, Lin et al. (2015b) proposed CTransR in which the entity pairs are clustered into different groups and where the pairs in the same group share the same relation vector. Xiao et al. (2016) developed TransG, a generative Bayesian non-parametric infinite mixture embedding model, to handle multiple relation semantics of an entity pair. Further, Wang et al. (2019) integrated logic rules into a translation based knowledge graph embedding model. Their method automatically mined logic rules from triples in a knowledge graph.

Previous work has also introduced external knowledge to learn better knowledge embeddings. Lin et al. (2015a) and Luo et al. (2015) utilized relation paths and Guo et al. (2015) integrated additional semantic information and enforced the embedding space to be semantically smooth so that entities in the same semantic category were close to each other in the embedding space. Wang et al. (2014) used entity names and Wikipedia anchors to align the embeddings of entities and words in the same space. In our work, we focus on converting knowledge relations from different knowledge sources to relations that are tailored to the NLI task. We then use this knowledge to illustrate the impact that both knowledge content and representation have on model performance.

2.3 Language Model Challenges

Pre-trained language models face several challenges and previous work has analyzed and illustrated their strenghts and weaknesses. Ettinger (2020) constructed a series of tests for language models and applied these to BERT to study strengths and weakness. Kassner and Schütze

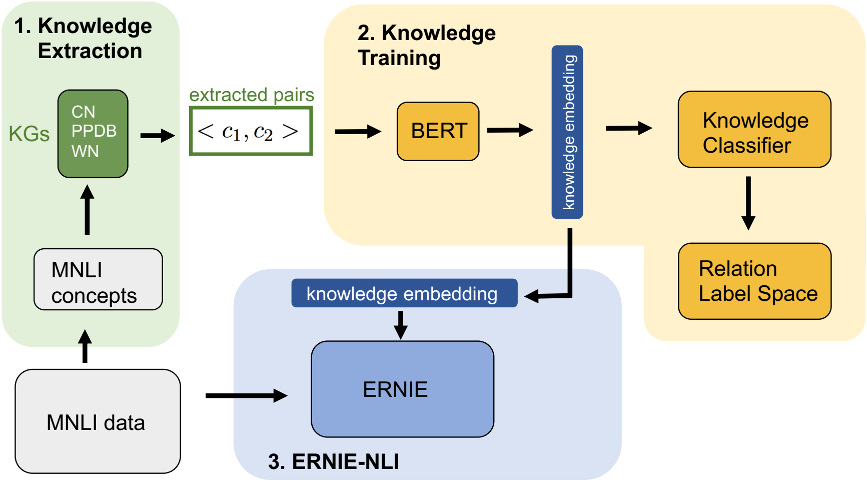

Figure 1: Components of the setup: (1) Knowledge Extraction Phase: Extracts knowledge content from external knowledge sources; (2) Knowledge Training Phase: Learns knowledge embeddings adapted to the NLI task; and (3) ERNIE-NLI: Trains NLI model with the integration of our learned knowledge embeddings.

(2020) added a component that focused on negation to the LAMA (LAnguage Model Analysis) evaluation framework (Petroni et al., 2019), showing that BERT failed on most negated statements. Talmor et al. (2019) designed eight reasoning tasks and illustrated that reasoning abilities are strongly context-dependent. Specific to NLI, Richardson et al. (2019) constructed challenging NLI datasets with new semantic fragments and showed that language models, though trained on NLI benchmark datasets, did not perform well on the new fragments. This previous work has shown that when applying pre-trained language models to a new task, a new domain, or new data variations, these models do not always perform well and additional knowledge may be needed to guide them. We examine how different types of knowledge impact language model decisions with respect to both content and polarity.

3 NLI corpus and External Knowledge

In this section, we introduce the particular NLI corpus and external knowledge sources used throughout this work.

3.1 NLI Corpus

MNLI, the Multi-Genre Natural Language Inference Corpus (Williams et al., 2018), consists of 433k sentence pairs annotated with entailment, contradiction, and neutral labels. The corpus covers various genres of both spoken and written text, and offers a wide range of style, various degrees of formality, and a diverse variety of topics and domains. This dataset is evaluated using standard accuracy.

3.2 External Knowledge Sources

We use several external knowledge sources to learn the relationships between concepts in our task.

ConceptNet (Speer et al., 2017) is a large semantic graph consisting of general knowledge. Concepts are related through predicates such as *IsA*(jazz, genre of music) and *AtLocation*(jazz, new orleans).

PPDB, Paraphrase Database (Ganitkevitch et al., 2013), contains over 220 million paraphrase pairs extracted from bilingual parallel corpora. Each paraphrase pair consists of two concepts that have a similar meaning.

WordNet (Miller, 1995) groups nouns, verbs, adjectives and adverbs into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are linked by different relations including synonym, antonymy, hypernymy, hyponymy, etc.

4 Methods

We introduce our terminology in Section 4.1. Then, we introduce the three steps of ERNIE-NLI: (1) knowledge extraction phase (content): extracting knowledge content from external knowledge sources (Section 4.2), (2) knowledge training phase (polarity): learning knowledge embeddings adapted to the NLI task (Section 4.3), and (3) NLI training phase: training our NLI model with the integration of learned knowledge embeddings (Section 4.4). The three phases are shown in Fig. 1.

4.1 Terminology

We use the following terms throughout the paper. For clarity, we will demonstrate each term given

(A)	*Premise:* I had an additional reason for that belief in the fact that all the cups found contained **sugar**, which Mademoiselle Cynthia never took in her coffee. *Hypothesis:* Mademoiselle Cynthia often took milk or **cream** in her **coffee**. *Label:* neutral *External Knowledge Pair: RelatedTo*(sugar, cream), *AtLocation*(sugar, coffee) *NLI Knowledge Pair:* `pos`(sugar, cream), `pos`(sugar, coffee)
(B)	*Premise:* Lalley also is enthused about other bar **efforts** on behalf of the poor, most notably the Legal Assistance Center will operate out of the new courthouse. *Hypothesis:* Lalley is enthusiastic about the bar's **initiative** to help the poor. *Label:* entailment *External Knowledge Pair: ReverseEntailment*(efforts, initiative) *NLI Knowledge Pair:* `pos`(efforts, initiative)

Table 1: NLI & Knowledge Pair Example.

the example in Table 1, Example (A).

External knowledge pair refers to a pair of two concepts from external knowledge sources, connected by an external knowledge relation, for example *RelatedTo*(sugar, cream). Each concept may be either a single word or a phrase.

External knowledge relation is the relation of the external knowledge pair. Each external knowledge source has a unique set of external knowledge relations. *RelatedTo* is an example of such a relation.

NLI knowledge pair refers to a pair of two concepts from NLI corpus, connected by an NLI knowledge relation, e.g., `pos`(sugar, cream).

NLI knowledge relation is the relation of the NLI knowledge pair. We define two NLI knowledge relations in Section 4.2: `pos`() and `neg`().

NLI pair refers to a pair of sentences, in which one sentence is the premise and the other is the hypothesis, as depicted in Table 1.

NLI label is entailment/neutral/contradiction.

4.2 NLI Knowledge Extraction

To represent external knowledge relations from different sources in a unified way, we define two NLI knowledge relations: `pos`() and `neg`(). A rule-based heuristic is developed to map the external knowledge relations to NLI knowledge relations. For example, in Table 1, we see that *RelatedTo* is mapped to `pos`(). Additionally, an external knowledge relation such as *Antonym* would be mapped to `neg`(). Each external knowledge relation is mapped to one NLI knowledge relation, where different external knowledge relations may be mapped to the same NLI knowledge relation. The specific mappings are listed in the appendix.

NLI knowledge pairs are extracted from each NLI pair. For the i-th NLI pair, with premise P and hypothesis H, we first identify all the concepts (single word or key phrase) in P and H using Python Keyphrase Extraction (PKE) (Boudin, 2016). We then extract each NLI knowledge pair $y(c_i^1, c_i^2)$ where $c_i^1 \subseteq P$ (a concept in the premise), $c_i^2 \subseteq H$ (a concept in the hypothesis) and where there exists an NLI knowledge relation y between c_i^1 and c_i^2. Considering Example (A) in Table 1, we see that c_i^1 = 'sugar', c_i^2 = 'cream', and y = `pos`().

There may be multiple NLI knowledge pairs in the i-th NLI pair of premise and hypothesis.

4.3 NLI Knowledge Learning

To learn the NLI knowledge embeddings, we add two additional components to BERT (Devlin et al., 2019). Thus, we learn the embedding of $y\{c_i^1, c_i^2\}$ in the following way. First, the sequence of knowledge tokens {[CLS] c_i^1 [SEP] c_i^2 [SEP]} is passed as input to BERT. Then, we take the subsequent contextual representations from BERT and pass them through a knowledge embedding layer (a linear layer) which casts our BERT representations into a knowledge embedding.

$$\mathbf{o} = \text{BERT}(c_i^1, c_i^2) \tag{1}$$
$$k_i = W_k(\mathbf{o}) + b_k \tag{2}$$

where $\mathbf{o}$ is the contextual representation from BERT, W_k and b_k are weights and bias of the knowledge embedding layer, and k_i is the knowledge embedding. Next, the knowledge embedding k_i is fed into the NLI knowledge relation classification layer for knowledge fine-tuning:

$$l_c = W_c(k_i) + b_c \tag{3}$$

$$y = \text{softmax}(l_c) \tag{4}$$

where W_c and b_c are weights and bias of the classification layer, and y is the NLI knowledge relation prediction. We use cross-entropy loss during training. In this way, we get the knowledge embedding associated with the NLI knowledge relation.

We learn the embeddings for all the NLI knowledge pairs in the i-th NLI pair in the training set such that we have a set of knowledge $K_i = \{k_i^1, \dots, k_i^m\}$ where m is the length of the knowledge sequence for the i-th NLI pair. We use these embeddings to enhance NLI training described in the next section. The knowledge embeddings are fixed during NLI training. Note that at inference time, we calculate the knowledge embedding of the relation between any two concepts in the premise and hypothesis via Equations 1 and 2, even if the two concepts are not included in the training set. This enables the model to handle unseen concepts and NLI knowledge relations in the inference data.

4.4 NLI Knowledge Enhanced NLI

We propose ERNIE-NLI, built on the ERNIE architecture (Zhang et al., 2019a), to integrate the knowledge embeddings learned in Section 4.3 into the NLI model.

4.4.1 ERNIE

ERNIE (Zhang et al., 2019a) was developed mainly for integrating knowledge graph information into the entity typing and relation extraction tasks. It has two stacked modules: (a) a textual encoder to capture token embeddings and (b) a knowledge encoder to inject the token-oriented knowledge into the textual encoder output. The textual encoder is a multi-layer bidirectional Transformer encoder, similar to BERT. The knowledge encoder concatenates the token embeddings (output from the textual encoder) and entity embeddings (pre-trained TransE embedding).

ERNIE defines two inputs to the model, a token sequence $T = \{w_1, \dots, w_n\}$ where n is the length of the token sequence, and a entity sequence that aligns to the given tokens as $E = \{e_1, \dots, e_m\}$ where m is the length of the entity sequence. ERNIE is then defined as:

$$\mathbf{u} = \text{ERNIE}(T, E) \tag{5}$$

For example, consider the following sentence:

Bob Dylan wrote *Blowin' in the Wind.*

To recognize the relation between Bob Dylan and Blowin' in the Wind, ERNIE concatenates the entity embeddings of Bob Dylan and Blowin' in the Wind with the corresponding token embeddings. For more details, please refer to the original paper (Zhang et al., 2019a).

4.4.2 ERNIE-NLI

Though ERNIE is mainly designed for the entity typing and relation extraction tasks, it also reports performance on the MNLI dataset. ERNIE does not show an improvement over BERT, even though it uses the information from the knowledge graph. We speculate that this is because the knowledge type (named entities) is neither the type of knowledge required for the NLI task nor domain-specific to the NLI task. In contrast to ERNIE, which directly uses TransE embeddings (which are not adapted to the NLI task), we propose ERNIE-NLI which uses knowledge embeddings trained on the NLI dataset and tailored for the NLI task.

Similar to ERNIE, two inputs are fed into ERNIE-NLI: a token sequence $T = \{w_1, \dots, w_n\}$ and a knowledge sequence, aligned to the given tokens, as $K = \{k_1, \dots, k_m\}$ where m is the length of the knowledge sequence. In contrast to ERNIE, knowledge relations are tailored to the NLI task and knowledge embeddings are trained on the NLI training data. Thus, our model definition becomes:

$$\mathbf{u} = \text{ERNIE}(T, K) \tag{6}$$

where our knowledge embeddings for K are fixed during NLI training, similar to the original setup. However, unlike the original setup, our knowledge embeddings are now adapted to the NLI task.

5 Experiment Setup

As introduced in Section 3, we examine various external knowledge sources. We describe the setups used in this work, all of which are combinations of these sources. The performance of each setup is reported in Section 6.

PC is the basic setup and includes Paraphrase Database (PPDB) and ConceptNet. In this setup, we find that the number of positive NLI knowledge relations is greater than the number of negative NLI knowledge relations. Thus, we design additional setups to balance the ratio of positive and negative relations.

PC&Bal balances the positive and negative NLI knowledge relations to 50%-50% by downsampling positive relations.

PCW adds negative NLI knowledge relations from WordNet to PC.
PCW&Bal balances the positive and negative NLI knowledge relations to 50%-50% on PCW by downsampling positive relations.

6 Results and Analysis

6.1 BERT Error Analysis

Before designing our experiments, we manually analyzed BERT misclassifiations on MNLI, which inspired the decisions regarding content and polarity of knowledge required for improved reasoning and performance. We achieved 83.90% on the MNLI dev set with BERT. We analyzed 40 misclassifications per MNLI domain, and found that across all domains, at least 50% of misclassifications required external knowledge to be resolved. We also found that the combination of ConceptNet and PPDB covered at least 70% of the required concepts for these misclassifications across all domains. Thus, we decided to investigate the impact of external knowledge on NLI models.

6.2 ERNIE-NLI Performance

We run both ERNIE and ERNIE-NLI on the MNLI corpus using our experimental setups. With respect to ERNIE as the baseline, the accuracy changes of ERNIE-NLI are shown in Table 2. As introduced in Section 5, PC&Bal has less positive relations than PC. We can see that in Table 2, PC has better performance on the entailment class than PC&Bal, but has worse performance on neutral and contradiction. Similarly, PCW achieves better performance on entailment than PCW&Bal and worse performance on neutral and contradiction.

PCW has more negative NLI knowledge relations than PC since PCW has additional negative relations from WordNet. As shown in Table 2, PC achieves better performance on the entailment class than PCW and worse performance on the neutral class. Similarly, PC&Bal has better performance on the entailment class than PCW&Bal and worse performance on neutral and contradiction classes.

These results demonstrate a correlation between knowledge polarity and NLI performance, specifically that adding positive knowledge can train an NLI model that is better at making entailment predictions, and that adding negative knowledge can train an NLI model that is better at making neutral and contradiction predictions. As shown in Table 2, the best setup for the entailment class is PC and the best setup for the contradiction and neutral classes is PCW&Bal. The accuracy of the two setups per label and on all labels are included in Table 3 below. Note that in both setups, ERNIE-NLI not only achieves better performance on the particular NLI class, but also achieves better total performance. While ERNIE-NLI achieves better performance in this knowledge-integration setup, for comparison we would like to point out that the state-of-the-art is achieved by T5-11B (Raffel et al., 2020), which achieves 92.2% on the MNLI test set.

Setup	Contra	Neutral	Entail
PC	-0.62	0.13	2.59
PC&Bal	0.22	0.96	-1.06
PCW	-1.00	0.66	1.41
PCW&Bal	0.59	1.47	-0.84

Table 2: ERNIE-NLI improvement over ERNIE in % Accuracy per Contradiction/Neutral/Entailment label.

Model	Contr.	Neut.	Ent.	Total
ERNIE	85.91	83.74	80.84	83.42
ERNIE-NLI E	85.29	83.87	**83.43**	**84.18**
ERNIE-NLI C&N	**86.50**	**85.21**	80.00	83.74

Table 3: % Accuracy per label for ERNIE and ERNIE-NLI using best setup for each label.

6.3 Classification Change Analysis

We further analyze the new errors per label made by ERNIE-NLI compared to ERNIE. Table 4 shows the number of **error** changes grouped by NLI labels, and demonstrates that all the increased error changes from ERNIE to ERNIE-NLI enhanced with PC (i.e., positive numbers in the row of PC) are false entailment classifications. This observation is consistent with the findings in Table 2: with the introduction of more positive than negative knowledge, our model becomes biased towards entailment. Similarly, all of the increased errors changes from ERNIE to ERNIE-NLI enhanced with PCW&Bal (i.e., positive numbers in the row of PCW&Bal) are false neutral predictions. More interestingly, in this PCW&Bal setup where the positive and negative knowledge is balanced, the new errors only occur when the gold label is entailment and all other errors decrease. These results indicate that the model is able to utilize knowledge in a way that reflects an understanding of the NLI label. When the knowledge is balanced,

Gold	Contra		Neutral		Entail	
Prediction	N	E	C	E	C	N
PC	-2	22	-26	24	-9	-81
PCW&Bal	0	-16	-20	-20	-5	117

Table 4: ERNIE-NLI error changes with respect to ERNIE. A positive value indicates that ERNIE-NLI makes more errors than ERNIE on that label and vice versa.

	Contra	Neutral	Entail	Total
0%	86.35	83.93	80.12	83.37
25%	86.25	83.80	80.52	83.44
50%	86.50	85.21	80.00	83.74
78%	84.91	84.40	82.25	83.80
100%	85.29	83.87	83.43	84.18

Table 5: ERNIE-NLI performance with respect to the portion of positive knowledge used during knowledge training.

the model better understands the boundary between entailment and contradiction.

To better understand knowledge effect on ERNIE-NLI, we conduct a series of experiments to answer the following questions:

- Is more knowledge better?
- How does knowledge polarity affect NLI classification?
- How is performance affected if there is new knowledge at inference time?

6.4 Knowledge Portion during Training

To investigate performance gains with respect to the addition of NLI knowledge, we report the NLI performance depending on the portion of positive knowledge used during NLI knowledge learning under the PC setup in Table 5, which shows how the incremental addition of positive knowledge during knowledge embedding training increases the NLI performance for the entailment label. Note that the total accuracy is increased as more positive knowledge is added.

6.5 Knowledge Type during Inference

An NLI contradiction pair may extract positive NLI knowledge relations and an entailment pair may extract negative NLI knowledge relations. We analyze the correlation between the presence of NLI knowledge relations and the prediction results on the dev set. Specifically, we compare the prediction changes from ERNIE to ERNIE-NLI using the PC setup. Table 6 shows these prediction changes. X → Y represents the NLI pairs where baseline ERNIE predicts X while ERNIE-NLI predicts Y. We also include the number of correct prediction changes (i.e., where Y is gold).

Label	Pos Rels		Neg Rels		None Rels	
C/N→E	160	(101)	22	(11)	138	(88)
C/E →N	156	(96)	24	(16)	140	(57)
N/E →C	129	(60)	22	(9)	82	(35)

Table 6: ERNIE-NLI classification changes with respect to ERNIE depending on presence of knowledge at inference time. Numbers without parenthesis are the total changes and numbers in the parenthesis are the correct changes.

Since we show results on the PC setup, we focus on the first row and first column in the table. The results in the first row indicate that a correct entailment classification with the presence of positive knowledge is observed to occur more often than with the presence of negative knowledge. The results in the first column indicate that a correct entailment classification with the presence of positive knowledge is observed to occur more often than a correct neutral or contradiction classification with positive knowledge. Thus, we see a strong correlation between the presence of positive knowledge and a correct entailment classification. This is a result of using the PC setup in this analysis, which is tailored for positive relations. Thus, while the correct entailment classification has the strongest correlation, we also see the strong effect of positive relations across all categories.

We would like to note that these findings are not discovered solely by looking at the label accuracies, as other classification shifts in this setting occur. We believe carrying out careful analyses, such as these, enable us to gain a deeper understanding of how knowledge affects the neural model, as we see clear trends in the effect of knowledge presence by polarity via this analysis.

6.6 Unseen Knowledge during Inference

To investigate our model's robustness in a common scenario where there are unseen knowledge relations in the evaluation data, we experiment with using only four external knowledge relations as NLI

Mapping	Contra	Neutral	Entail
Constrained	0.09	0.48	-0.17
Unconstrained	-0.31	0.57	0.63

Table 7: ERNIE-NLI % Accuracy changes for handling unseen relations with respect to ERNIE.

knowledge relations during training. The four relations are: RelatedTo, IsA, Independent, Antonym. During inference, we design two scenarios.

First, we design a constrained scenario in which new relations during inference time are dropped. For example, if an “Entails" relation exists between two concepts according to the knowledge sources, the knowledge is discarded, since it is not included in one of the four relations.

Second, we design an unconstrained scenario that computes the knowledge embedding at inference time. The sequence of the two concepts linked by the “Entails" relation, {[CLS] c_i^1 [SEP] c_i^2 [SEP]}, are fed into the BERT layer in Equation (1) and knowledge embedding layer in Equation (2) to get the knowledge embedding.

We compare the performance of the two scenarios in Table 7. The unconstrained scenario performs better than the constrained scenario, especially on the entailment label, given that there is more positive knowledge. The result shows ERNIE-NLI’s capability of utilizing unseen knowledge relations to improve NLI, indicating the robustness of ERNIE-NLI in providing good predictions even if the inference data has shifted.

7 Examples

In this section, we discuss the two examples depicted in Table 1, to show how external knowledge can assist models on the NLI task.

7.1 Introducing World Knowledge

Integrating external knowledge can equip the model with world knowledge it did not have access to before. In Table 1, Example (A), the baseline model without external knowledge predicts contradiction, which is incorrect. Our ERNIE-NLI model with external knowledge predicts neutral, which is correct. The external knowledge used in this example is *RelatedTo*(sugar, cream) and *AtLocation*(sugar, coffee). The baseline model seems to predict this as contradiction mainly because the premise states *never ... in her coffee* while the hypothesis states *in her coffee*. The external knowledge helps correctly align the components: *sugar* and *cream*. Note that although the external knowledge indicates that *sugar* is related to *cream*, it does not necessarily yield an entailment prediction as the context is still being taking into consideration by the model, which understands that *sugar* is the main condition for entailment and that *cream* and *sugar* are not synonymous in this context.

7.2 Emphasizing Phrase Similarity

The model looks for similar words or phrases when it judges whether the hypothesis can be entailed from the premise. In the baseline model, the contextual embeddings alone are not strong enough to drive the prediction. In Table 1, Example (B), the baseline prediction is contradiction, which is wrong. Our ERNIE-NLI model with external knowledge predicts entailment, which is correct. The key knowledge required for this example is *Paraphrase*(efforts, initiative). By adding this paraphrase knowledge, the enhanced model recognizes the entailment relation of the pair.

8 Conclusion

We propose ERNIE-NLI, an NLI model that integrates external knowledge to enhance NLI performance. Our external knowledge representations are tailored to the NLI task and trained to adapt to NLI data requirements. We show that our model enhanced with external knowledge achieves better performance than the previous ERNIE model with non-adapted knowledge depending on the knowledge utilized. We examine these results with several analysis experiments to enable strong conclusions about the correlation between knowledge and NLI classification. Results also demonstrate that the model is able to handle unseen knowledge when the inference data shifts from training data.

Acknowledgments

We thank the reviewers for their useful feedback. This work was performed while Lisa interned at Bloomberg, and was later supported by DARPA MCS Grant N66001-19-2-4031, NSF-CAREER Award 1846185, and an NSF PhD Fellowship. The views are those of the authors and not of the funding agency.

References

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NeurIPS*.

Johan Bos and Katja Markert. 2005. Recognising textual entailment with logical inference. In *HLT-EMNLP*.

Florian Boudin. 2016. pke: an open source python-based keyphrase extraction toolkit. In *COLING*.

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

Samuel R. Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D. Manning, and Christopher Potts. 2016. A fast unified model for parsing and sentence understanding. In *ACL*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2018. Neural natural language inference models enhanced with external knowledge. In *ACL*.

Jihun Choi, Kang Min Yoo, and Sang-goo Lee. 2018. Learning to compose task-specific tree structures. In *AAAI*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *NAACL-HLT*, pages 758–764.

Shu Guo, Quan Wang, Bin Wang, Lihong Wang, and Li Guo. 2015. Semantically smooth knowledge graph embedding. In *ACL*, pages 84–94, Beijing, China. Association for Computational Linguistics.

Adrian Iftene and Alexandra Balahur. 2007. Hypothesis transformation and semantic variability rules used in recognizing textual entailment. In *ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*.

Jinbae Im and Sungzoon Cho. 2017. Distance-based self-attention network for natural language inference. *arXiv preprint arXiv:1712.02047*.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. *ACL*.

Seonhoon Kim, Inho Kang, and Nojun Kwak. 2019. Semantic sentence matching with densely-connected recurrent and co-attentive information. In *AAAI*.

Anne Lauscher, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2020. Specializing unsupervised pretraining models for word-level semantic similarity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1371–1383.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. Sensebert: Driving some sense into bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667.

Yankai Lin, Zhiyuan Liu, Huanbo Luan, Maosong Sun, Siwei Rao, and Song Liu. 2015a. Modeling relation paths for representation learning of knowledge bases. In *EMNLP*, pages 705–714, Lisbon, Portugal. Association for Computational Linguistics.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015b. Learning entity and relation embeddings for knowledge graph completion. In *AAAI*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *ACL*.

Yuanfei Luo, Quan Wang, Bin Wang, and Li Guo. 2015. Context-dependent knowledge graph embedding. In *EMNLP*, pages 1656–1661, Lisbon, Portugal. Association for Computational Linguistics.

Bill MacCartney and Christopher D Manning. 2009. An extended model of natural logic. In *IWCS*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2015. Natural language inference by tree-based convolution and heuristic matching. In *ACL*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Matthew E Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language*

Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 43–54.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? In *EMNLP*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Kyle Richardson, Hai Hu, Lawrence S Moss, and Ashish Sabharwal. 2019. Probing natural language inference models through semantic fragments. *arXiv preprint arXiv:1909.07521*.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Sen Wang, and Chengqi Zhang. 2018. Reinforced self-attention network: a hybrid of hard and soft attention for sequence modeling. In *IJCAI*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *ACL*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. *AAAI*.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2019. olmpics–on what language model pre-training captures. *arXiv preprint arXiv:1912.13283*.

Pengwei Wang, Dejing Dou, Fangzhao Wu, Nisansa de Silva, and Lianwen Jin. 2019. Logic rules powered knowledge graph embedding. *arXiv preprint arXiv:1903.03772*.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph and text jointly embedding. In *EMNLP*.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. In *IJCAI*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL-HLT*.

Han Xiao, Minlie Huang, and Xiaoyan Zhu. 2016. Transg: A generative model for knowledge graph embedding. In *ACL*, pages 2316–2325.

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

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019a. Ernie: Enhanced language representation with informative entities. In *ACL*.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2019b. Semantics-aware bert for language understanding. *arXiv preprint arXiv:1909.02209*.

A Appendix

A.1 Knowledge Mapping

Table 8 shows the external knowledge relations that are mapped to positive and negative NLI knowledge relations.

A.2 Hyperparameter Settings

For our experiments, we did not tune hyperparameters but rather selected our settings to be consistent with Zhang et al. (2019a). We used batch size 12, learning rate 2e-5, and random seed 42. We did 1 epoch of relation training and 4 epochs of NLI training. We hold these settings constant across all experiments.We built on the framework released by Zhang et al. (2019a), which included a pytorch implementation of ERNIE, and used all versions and infrastructures included in their implementation.

Course Grained	**Fine-Grained**
Negative	Antonym
	DistinctFrom
	Exclusion
	Unrelated
Positive	IsA
	Synonym
	RelatedTo
	HasFirstSubevent
	MannerOf
	NotCapableOf
	CausesDesire
	MotivatedByGoal
	HasProperty
	Entails
	ForwardEntailment
	CreatedBy
	Equivalence
	DerivedFrom
	dbpedia
	OtherRelated
	Unrelated
	MadeOf
	Desires
	ReceivesAction
	SimilarTo
	EtymologicallyRelatedTo
	HasLastSubevent
	NotHasProperty
	HasSubevent
	DefinedAs
	CausesDesire
	AtLocation
	HasA
	Independent
	ReverseEntailment
	FormOf
	HasContext
	InstanceOf
	PartOf
	NotDesires
	HasPrerequisite
	UsedFor
	CapableOf

Table 8: Fine-grained to course-grained mapping for External Knowledge Relations to NLI Knowledge Relations.