Paraphrase Identification via Textual Inference

Ning Shi **Bradley Hauer** **Jai Riley** **Grzegorz Kondrak**
Alberta Machine Intelligence Institute (Amii)
Department of Computing Science
University of Alberta, Edmonton, Canada
{ning.shi,bmhauer,jrbuhr,gkondrak}@ualberta.ca

Abstract

Paraphrase identification (PI) and natural language inference (NLI) are two important tasks in natural language processing. Despite their distinct objectives, an underlying connection exists, which has been notably under-explored in empirical investigations. We formalize the relationship between these semantic tasks and introduce a method for solving PI using an NLI system, including the adaptation of PI datasets for fine-tuning NLI models. Through extensive evaluations on six PI benchmarks, across both zero-shot and fine-tuned settings, we showcase the efficacy of NLI models for PI through our proposed reduction. Remarkably, our fine-tuning procedure enables NLI models to outperform dedicated PI models on PI datasets. In addition, our findings provide insights into the limitations of current PI benchmarks.[1]

1 Introduction

Semantic relationships have been the subject of extensive research, and play pivotal roles in natural language processing (Burdick et al., 2022; Hauer and Kondrak, 2023; Pàmies et al., 2023; Peng et al., 2023a; Wahle et al., 2023), including the study and evaluation of the reasoning capabilities of language models (Liu et al., 2019; Yang et al., 2019). Two important tasks that depend on semantic relations between sentences are paraphrase identification (PI; Bai et al., 2023; Peng et al., 2023b) (NLI; Williams et al., 2018; Nie et al., 2020; Williams et al., 2022). PI is the task of deciding whether two sentences are in the paraphrase relation, that is, whether they convey the same meaning (Bhagat and Hovy, 2013). NLI involves three labels that describe the relationship between two sentences: entailment, contradiction, and neutral (MacCartney, 2009).

Our focus is specifically on detecting textual entailment, as indicated by the first of these categories (Bos and Markert, 2005; Dagan et al., 2005; Poliak, 2020), or, more generally, textual inference (Manning, 2006), which is the relation between sentences where one can be inferred from the other in a given context. Take the example from SNLI (Bowman et al., 2015); while the premise "this man is surfing" does not always entail the hypothesis "a man is on water", the broader context may make it clear that the word *surfing* refers to an aquatic activity rather than website browsing, and so the latter sentence can be inferred from the former.

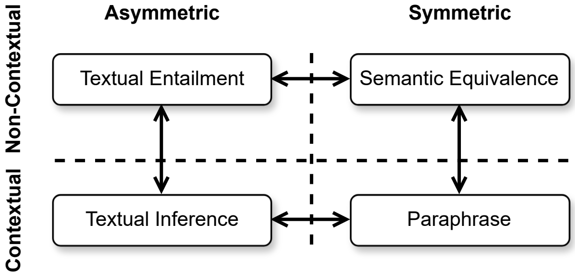

Figure 1: Four sentence-level relations in terms of symmetry and contextuality. Arrows indicate interdependence between the relations (Section 2).

Prior work has hypothesized that paraphrasing corresponds to bidirectional textual entailment; see, for example, the surveys of Androutsopoulos and Malakasiotis (2010) and Madnani and Dorr (2010). However, to the best of our knowledge, the only work that empirically investigates the connection between these two tasks is Seethamol and Manju (2017). They incorporate a blend of modules, including word sense disambiguation for sentence similarity and a Markov logic network for probabilistic inference, which complicates the analysis of the interplay between paraphrases and entailment. Moreover, their approach aligns more with traditional PI methods than with our approach, and lacks any theoretical formalization.

In this work, we formalize prior informal observations on the relationship between textual entail-

[1] We make our code and data publicly available on GitHub: https://github.com/ShiningLab/PI2NLI.

ment and paraphrasing into a coherent theoretical framework (Figure 1). We formally define four semantic relations and classify them according to two criteria: symmetry and contextuality. This formalization implies a practical reduction of PI to NLI, which we empirically validate by employing two widely used pre-trained transformer-based language models, RoBERTa and XLNet. We introduce a dataset adaptation process for fine-tuning an NLI model for PI, and test our implementation on six PI benchmarks. Our results indicate that in the fine-tuned setting, our PI to NLI reduction can actually yield better performance compared to the direct application of a PI system. This provides strong support for the utility of our reduction, and the theoretical model upon which it is based.

2 Methodology

In this section, we present our theoretical framework linking four semantic relations. We also introduce a novel method for fine-tuning an NLI model for PI, proposing a dataset adaptation procedure that converts PI datasets to labeled NLI instances.

2.1 Equivalence and Paraphrasing

We define the *semantic equivalence relation* (SEQ) as follows:

SEQ(S_1, S_2) := "the sentences S_1 and S_2 convey the same meaning"

The *paraphrase relation* (PR) between sentences is related to semantic equivalence; specifically, SEQ implies PR. Our definition of PR is *contextual*, so that it also admits semantic equivalence in a broader context, which may include common sense and world knowledge.

PR(C, S_1, S_2) := "the sentences S_1 and S_2 convey the same meaning given the context C"

Bhagat and Hovy (2013) refer to this type of paraphrases as *quasi-paraphrases*; for example:

- S_1: *We must work hard to win this election.*
- S_2: *The Democrats must work hard to win this election.*

We postulate the following relationship between the semantic equivalence and paraphrase relations:

$$\mathrm{SEQ}(S_1, S_2) \Leftrightarrow \forall C : \mathrm{PR}(C, S_1, S_2)$$

2.2 Entailment and Inference

Textual entailment (TE) is a directional relation between sentences which holds if the truth of one sentence follows from another sentence (Dagan and Glickman, 2004):

TE(S_1, S_2) := "the sentence S_2 can be inferred from the sentence S_1"

The proposition that T entails H is denoted as $T \models H$. The entailment relation is not symmetric: $T \models H$ does *not* imply $H \models T$.

Following prior work, we assume that sentences are semantically equivalent if and only if each entails the other:

$$\mathrm{SEQ}(S_1, S_2) \Leftrightarrow \mathrm{TE}(S_1, S_2) \wedge \mathrm{TE}(S_2, S_1)$$

Finally, we define *textual inference* (TI) as a contextual generalization of textual entailment which takes into account the broad context of the statements, which may include common sense and world knowledge (Manning, 2006):

TI(C, S_1, S_2) := "the sentence S_2 can be inferred from the sentence S_1 given the context C"

Intuitively, $TI(C, S_1, S_2)$ expresses the following inference property: $(C + S_1) \models S_2$.

Analogous to the relationship between SEQ and PR, we postulate the following relationship between TE and TI:

$$\mathrm{TE}(S_1, S_2) \Leftrightarrow \forall C : \mathrm{TI}(C, S_1, S_2)$$

The following proposition establishes a connection between PR and TI:

Proposition 1 *Given context C, sentences S_1 and S_2 are paraphrases if and only if they can be mutually inferred from each other.*

$$\mathrm{PR}(C, S_1, S_2) \Leftrightarrow \mathrm{TI}(C, S_1, S_2) \wedge \mathrm{TI}(C, S_2, S_1)$$

Thus, the paraphrase relation can be viewed as the conjunction of the inference relation in both directions.

2.3 Dataset Adaptation

Building on our theoretical formalization, we posit that the task of PI, which depends on detecting the PR relation, can be reduced to NLI, specifically the detection of the TI relation. To implement and test our PI to NLI reduction – henceforth PI2NLI – we present a novel fine-tuning procedure that allows an NLI model to be fine-tuned for solving PI instances. Our goal is to mitigate biases stemming from the transfer learning and any domain-specific disparities or other properties of the data that may degrade performance on PI datasets. Our dataset adaptation procedure transforms PI datasets to be compatible with NLI systems so as to facilitate fine-tuning on adapted PI data.

We convert each positive PI instance into two distinct positive NLI instances, one in each direction, indicating mutual TI between two paraphrases, as postulated in Proposition 1. Conversely, since determining in which direction TI fails to hold in a negative PI instance is not straightforward, we generate a negative NLI instance in a random direction. While this heuristic is not theoretically justified, we found that it works well in practice.

3 Experiments

The experiments in this section are aimed at validating the proposed theoretical framework. Additional data specifics and training details can be found in Appendices B and C.

3.1 Models

We implement and test our reduction with each of two freely available transformer-based (Vaswani et al., 2017) language models, RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019). Specific model names have been provided in the footnotes. We choose them because of their low hardware requirements, and their status as well-known and well-studied models (Peng et al., 2022). The primary distinction between them lies in their design: RoBERTa is an autoencoding-based model, while XLNet is an autoregressive model. Note that the prior works we will mainly compare to are as recent as 2022, thus we gain no advantage from our choice of models.

In our implementation, we apply the NLI classification head because pre-trained NLI models are readily available (Nie et al., 2020). We consider the relation labeled as "entailment" in the NLI datasets as TI rather than TE because the positive instances typically require broader contextual knowledge, as exemplified by the "surfing" instance in Section 1. Since NLI models are not typically trained on paraphrase data (PI being an entirely separate task from NLI), this maintains a sound experimental setup.

Since recognizing TI is a binary task (outputs are positive or negative), while NLI is a ternary task (outputs are entailment, neutral, or contradiction), we require a means of converting labeled TI instances to NLI instances (so that we can fine-tune NLI models), and NLI outputs to TI outputs (so that we can evaluate them). We map positive TI labels to "entailment" NLI labels and negative TI labels to "neutral" or "contradiction" labels at random. We map "entailment" NLI output to a positive TI classification, and "neutral" or "contradiction" to a negative TI classification. Further details and discussion can be found in Appendix A.

Data	#Train.	#Valid.	#Test	Test Pos.%
PIT	11,530	4,142	838	20.88
QQP	384,290	10,000	10,000	50.00
MSRP	3,668	408	1,725	66.49
PAWS QQP	11,988	8,000	677	28.21
PAWS Wiki	49,401	8,000	8,000	44.20
PARADE	7,550	1,275	1,357	47.90

Table 1: Statistics of all six benchmarks, including the positive rate of the test set (Test Pos.%).

For the zero-shot application of PI2NLI, pi2nli$_{\text{zero}}$, we employ two trained NLI models: RoBERTa$_{\text{nli}}$[2] and XLNet$_{\text{nli}}$[3]. For the fine-tuned version, pi2nli, these models undergo fine-tuning on the NLI dataset derived from the corresponding PI dataset through dataset adaptation (Section 2.3). This yields a TI (or, more accurately, NLI) model adapted for PI following our PI2NLI reduction.

3.2 Setup

Data We test our reduction on six PI benchmarks: PIT (Xu et al., 2015), QQP (Iyer et al., 2017), MSRP (Dolan and Brockett, 2005), PAWS QQP (Zhang et al., 2019), PAWS Wiki (Zhang et al., 2019), and PARADE (He et al., 2020). We follow the data processing established by prior work (He et al., 2020; Peng et al., 2022). Detailed specifications of each dataset are provided in Table 1.

Baselines We adopt baselines from previous studies, citing each source for reference. Beyond referencing prior work, we set new benchmarks pi by training dedicated PI models using the same language models as pi2nli, alongside vanilla RoBERTa and XLNet.[4] Furthermore, we ensure that all classification heads are initialized from scratch. This facilitates a controlled comparison to isolate the distinct contributions of the PI2NLI reduction from the language models used. We meticulously follow the experimental setups and data preprocessing detailed in the referenced works, particularly aligning with the protocol established by Peng et al. (2022) for hyperparameter tuning.

[2] roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli

[3] xlnet-large-cased-snli_mnli_fever_anli_R1_R2_R3-nli

[4] roberta-large, xlnet-large-cased

Backbone	Method	PIT	QQP	MSRP	PAWS QQP	PAWS Wiki	PARADE
–	Random	27.18	50.31	56.47	35.01	46.94	51.22
BERT$_{\text{base}}$	Reimers and Gurevych (2019)	$52.03_{\pm 1.44}$	$90.78_{\pm 0.09}$	$81.67_{\pm 0.46}$	$66.01_{\pm 0.45}$	$81.57_{\pm 0.53}$	–
	Peng et al. (2021)	$59.11_{\pm 0.93}$	$90.41_{\pm 0.09}$	$81.70_{\pm 0.17}$	$66.22_{\pm 0.75}$	$81.14_{\pm 0.81}$	–
	Peng et al. (2022)	$59.19_{\pm 1.85}$	$90.74_{\pm 0.06}$	$83.42_{\pm 0.23}$	$68.85_{\pm 0.73}$	$82.60_{\pm 0.18}$	–
BERT$_{\text{large}}$	He et al. (2020)	74.60	87.70	89.30	–	93.30	70.90
RoBERTa$_{\text{base}}$	Reimers and Gurevych (2019)	$52.67_{\pm 2.75}$	$90.79_{\pm 0.09}$	$81.69_{\pm 0.53}$	$67.35_{\pm 0.97}$	$81.42_{\pm 0.93}$	–
	Peng et al. (2022)	$59.50_{\pm 2.74}$	$90.76_{\pm 0.03}$	$83.22_{\pm 0.46}$	$69.68_{\pm 0.72}$	$82.87_{\pm 0.35}$	–
RoBERTa$_{\text{large}}$	pi (Liu et al., 2019)	$81.20_{\pm 0.89}$	$91.66_{\pm 0.22}$	$91.17_{\pm 0.15}$	$88.92_{\pm 1.09}$	$\mathbf{94.05_{\pm 0.22}}$	$71.10_{\pm 7.18}$
XLNet$_{\text{large}}$	pi (Yang et al., 2019)	$56.39_{\pm 32.39}$	$73.19_{\pm 40.92}$	$87.51_{\pm 4.36}$	$89.83_{\pm 1.24}$	$74.91_{\pm 41.88}$	$59.02_{\pm 32.82}$
RoBERTa$_{\text{nli}}$	pi (Nie et al., 2020)	$79.64_{\pm 1.72}$	$91.62_{\pm 0.28}$	$91.48_{\pm 0.68}$	$\mathbf{90.06_{\pm 1.81}}$	$93.89_{\pm 0.22}$	$74.65_{\pm 0.64}$
	pi2nli$_{\text{zero}}$ (ours)	10.70	53.03	35.92	61.36	71.40	27.00
	pi2nli (ours)	$\mathbf{83.64_{\pm 1.44}}$	$\mathbf{92.27_{\pm 0.14}}$	$\mathbf{92.38_{\pm 0.30}}$	$88.67_{\pm 1.84}$	$93.87_{\pm 0.18}$	$\mathbf{75.04_{\pm 0.85}}$
XLNet$_{\text{nli}}$	pi (Nie et al., 2020)	$78.80_{\pm 0.82}$	$91.27_{\pm 0.30}$	$91.00_{\pm 0.63}$	$89.68_{\pm 0.38}$	$93.66_{\pm 0.24}$	$73.97_{\pm 0.21}$
	pi2nli$_{\text{zero}}$ (ours)	18.46	60.28	50.38	56.00	69.97	33.74
	pi2nli (ours)	$82.07_{\pm 1.31}$	$91.95_{\pm 0.20}$	$91.41_{\pm 0.40}$	$87.55_{\pm 1.26}$	$93.90_{\pm 0.35}$	$74.24_{\pm 0.75}$

Table 2: F1 scores (%) of PI2NLI in zero-shot (pi2nli$_{\text{zero}}$) and fine-tuned (pi2nli) settings, compared with the Random and pi baselines we implemented, as well as prior methods cited. Scores highlighted in bold signify the best performance with a p-value < 0.005, denoting high statistical significance.

Metrics To address the inherent class imbalance in most datasets and follow prior work (Peng et al., 2022), we use the F1 score as our primary evaluation metric. We run each method on each dataset five times, using each integer from 0 to 4 as a random seed, and report the average F1 score.

3.3 Results

We present our results in Table 2.

Zero-shot The zero-shot performance of PI2NLI is erratic, with highly variable F1 scores across datasets. Indeed, pi2nli$_{\text{zero}}$ outperforms the random baseline on only half of the datasets. Our analysis reveals that this is not indicative of a flaw in our PI2NLI reduction but rather due to inherent flaws in the PI benchmarks. Specifically, the annotations in these datasets do not strictly conform to the criteria imposed by our hypothesis. Table 3 highlights instances where paraphrasing-induced information loss disrupts mutual TI, leading to discrepancies between the original PI labels (Y_{PI}) and the outputs ($\hat{Y}_{\text{PI}}$) derived from the PI2NLI hypothesis. In essence, our results suggest that PI2NLI is able to identify and rectify inconsistencies in PI benchmarks. Such inconsistencies also suggest that context information essentially represents the dataset-specific distribution in practice: a paraphrase identified in one dataset might not necessarily be considered a valid paraphrase in the other. Taken together, these findings strongly suggests the need for a dataset adaptation procedure, to prepare the model for the unique properties of each dataset.

Fine-tuning Contrariwise, the fine-tuned version of our PI2NLI reduction yields consistently high F1 scores, outperforming the reported results obtained by prior work on all six datasets. In particular, the F1 score of the RoBERTa$_{\text{large}}$-based PI2NLI implementation increases from 10.70 to 83.64 on the PIT dataset. Notably, our top performances of 92.27 on QQP and 75.04 on PARADE also surpass the 89.6 (Peng et al., 2023b) and 74.06 (Bai et al., 2023) reported by the latest work respectively. This demonstrates that our dataset adaptation procedure successfully empowers NLI models to adapt to the peculiarities of various PI datasets and to yield state-of-the-art results. Moreover, our experiments show that PI2NLI consistently outperforms dedicated PI models using the same underlying language models on four of six datasets. This controlled experiment therefore confirms that the performance gains achieved can be attributed to our PI2NLI reduction, rather than other factors like the differing model capacities.

Pre-training Another critical observation is that pre-training[5] on additional NLI data leads to better and more stable fine-tuned performance on PI tasks. This observation is especially evident when transitioning pi from XLNet$_{\text{large}}$ to XLNet$_{\text{nli}}$. While it is a common belief that pre-training on additional tasks (e.g., NLI) could inherently improve performance on one certain task (e.g., PI), this is

[5] We regard "pre-training" as any foundational training conducted prior to our task-specific fine-tuning in this work.

Input	$S_1 \models S_2$	$S_2 \models S_1$	$\hat{Y}_{PI}$	Y_{PI}
S_1: The district also sent letters yesterday informing parents of the situation . S_2: Parents received letters informing them of the possible contamination yesterday .	T	T	T	T
S_1: Two kids from Michigan are in today 's third round . S_2: Both will compete in today 's third round , which is all oral examination .	F	F	F	F
S_1: Pacific Northwest has more than 800 employees , and Wells Fargo has 2,400 in Washington . S_2: It has 800 employees , compared with Wells Fargo 's 2,400 .	T	F	F	T
S_1: Six Democrats are vying to succeed Jacques and have qualified for the Feb. 3 primary ballot . S_2: Six Democrats and two Republicans are running for her seat and have qualified for the Feb. 3 primary ballot .	F	T	F	T

Table 3: Four PI instances that differ in the detected entailment direction. Although all eight individual TI outputs are arguably correct, the last two instances are counted as false negatives.

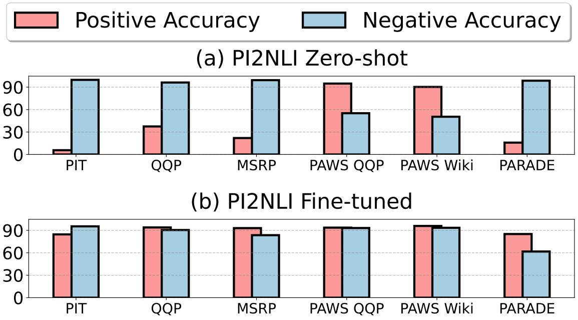

Figure 2: The results of (a) $\text{pi2nli}_{\text{zero}}$ and (b) pi2nli using $\text{RoBERTa}_{\text{nli}}$ in Table 2, separated into positive and negative accuracy.

not always a given. Several factors could potentially lead to a negative impact after such additional pre-training. These include domain mismatches, biases inherent in the pre-training data, and the phenomenon of catastrophic forgetting (McCloskey and Cohen, 1989). Following NLI pre-training, the improved performance of PI serves as a positive indicator. They support our hypothesis of a closely related and synergistic relationship between PI and NLI. This synergy is not automatic but is indicative of the effective transfer of relevant skills and knowledge from NLI to PI tasks.

Boundary In Figure 2, we split the results into positive and negative accuracy. In (a), $\text{pi2nli}_{\text{zero}}$ tends to have relatively higher negative accuracy, leading to a lower likelihood of classifying sentences as paraphrases. In (b), both positive and negative accuracy of pi2nli increase and become more balanced. This supports our earlier findings that, in order to perform better in the PI task, NLI models can correct their decision boundaries after fine-tuning. We view this adjustment as the process of how models learn the context inherent in each PI dataset.

PAWS Our error analysis reveals that the results of pi2nli on PAWS QQP and PAWS Wiki are due to the presence of adversarial examples (Zhang et al., 2019). This becomes particularly evident when comparing the QQP results with those of PAWS QQP, as both derive from the same source. These PAWS datasets are augmented with paraphrase adversaries to offer refined versions of the original datasets, presenting a challenge for models to predict the correct outcomes. Applying PI2NLI requires an NLI model to predict the TI relation in each direction. Therefore, the impact of the paraphrase adversaries becomes more apparent due to error accumulation from making two predictions.

4 Conclusion

We have presented a novel theoretical and empirical study of the relationship between two important semantic tasks, PI and NLI, a topic that has remained largely unexplored. Our experiments provide strong evidence that our innovative PI2NLI reduction, combined with fine-tuning on the NLI data facilitated by our dataset adaptation procedure, yields substantial F1 improvements on the PI task, outperforming dedicated PI models on benchmark PI datasets. The variable outcomes observed when applying PI2NLI in a zero-shot setting also offer insights into the existing limitations of the current PI datasets. In addition to advancing the state of the art, our findings offer valuable insights into the relation between PI and NLI, and set the stage for further investigation.

Limitations

While our work has made significant strides in understanding the four semantic relations, it is not without its limitations.

Firstly, our zero-shot results suggest mismatches between our theoretical proposition and existing

PI benchmarks. These benchmarks may not adequately capture the bidirectional inference relation integral to genuine paraphrase identification.

Secondly, our study focuses on the application of NLI models in solving PI tasks through the PI2NLI reduction, but there are still avenues left to explore. For instance, augmenting the PI dataset with an NLI one could potentially yield new insights.

Finally, our study has been NLI-centric so far, allowing us to delve deeply into the potential of NLI models in PI tasks. However, there is an opportunity for future research to explore the relationship from a PI-centric perspective. This could include investigating the capability of PI models in solving NLI tasks. A more balanced exploration would provide a more comprehensive understanding of the four semantic relations.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Alberta Machine Intelligence Institute (Amii).

References

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

Jun Bai, Chuantao Yin, Hanhua Hong, Jianfei Zhang, Chen Li, Yanmeng Wang, and Wenge Rong. 2023. Permutation invariant training for paraphrase identification. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Johan Bos and Katja Markert. 2005. Recognising textual entailment with logical inference. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 628–635, Vancouver, British Columbia, Canada.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal.

Laura Burdick, Jonathan K. Kummerfeld, and Rada Mihalcea. 2022. Using paraphrases to study properties of contextual embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4558–4568, Seattle, United States. Association for Computational Linguistics.

Ido Dagan and Oren Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability. *Learning Methods for Text Understanding and Mining*, 2004(26-29):2–5.

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

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Bradley Hauer and Grzegorz Kondrak. 2023. Taxonomy of problems in lexical semantics. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9833–9844, Toronto, Canada.

Yun He, Zhuoer Wang, Yin Zhang, Ruihong Huang, and James Caverlee. 2020. PARADE: A New Dataset for Paraphrase Identification Requiring Computer Science Domain Knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7572–7582, Online. Association for Computational Linguistics.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. Quora question pairs. *First Quora Dataset Release: Question Pairs*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Bill MacCartney. 2009. *Natural language inference*. Stanford University.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.

Christopher D Manning. 2006. Local textual inference: it's hard to circumscribe, but you know it when you see it–and nlp needs it. Technical report, Stanford University.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online.

Marc Pàmies, Joan Llop, Francesco Multari, Nicolau Duran-Silva, César Parra-Rojas, Aitor Gonzalez-Agirre, Francesco Alessandro Massucci, and Marta Villegas. 2023. A weakly supervised textual entailment approach to zero-shot text classification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 286–296, Dubrovnik, Croatia.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Qiwei Peng, David Weir, and Julie Weeds. 2021. Structure-aware sentence encoder in bert-based Siamese network. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 57–63, Online.

Qiwei Peng, David Weir, and Julie Weeds. 2023a. Testing paraphrase models on recognising sentence pairs at different degrees of semantic overlap. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 259–269, Toronto, Canada. Association for Computational Linguistics.

Qiwei Peng, David Weir, Julie Weeds, and Yekun Chai. 2022. Predicate-argument based bi-encoder for paraphrase identification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5579–5589, Dublin, Ireland.

Rui Peng, Zhiling Jin, and Yu Hong. 2023b. GBT: Generative boosting training approach for paraphrase identification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6094–6103, Singapore. Association for Computational Linguistics.

Adam Poliak. 2020. A survey on recognizing textual entailment as an NLP evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 92–109, Online.

Lutz Prechelt. 1998. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China.

S Seethamol and K Manju. 2017. Paraphrase identification using textual entailment recognition. In *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, pages 1071–1074. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Jan Philip Wahle, Bela Gipp, and Terry Ruas. 2023. Paraphrase types for generation and detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12148–12164, Singapore. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. 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)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Adina Williams, Tristan Thrush, and Douwe Kiela. 2022. ANLIzing the adversarial natural language inference dataset. In *Proceedings of the Society for Computation in Linguistics 2022*, pages 23–54, online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online.

Wei Xu, Chris Callison-Burch, and Bill Dolan. 2015. SemEval-2015 task 1: Paraphrase and semantic similarity in Twitter (PIT). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 1–11, Denver, Colorado.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. 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)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

A Dataset Adaptation

The alignment of PI data with NLI data starts with converting PI data to NLI format, as outlined in Section 2.3. While converting positive PI instances to positive NLI instances is straightforward, that for negative NLI instances is not. A negative PI instance is transformed into a negative NLI instance in one direction. When fine-tuning the NLI model, both "contradiction" and "neutral" are used to represent these negative NLI instances. In this context, a FALSE label is randomly assigned as either "contradiction" or "neutral" in NLI. This is justified in the context of our work because both labels can align with a negative TI relation.

Determining the precise TI direction and corresponding NLI class without additional resources or explicit human judgment presents a significant challenge. Hence, we adopted random sampling as a practical solution in our research. However, we recognize that further refining this aspect, such as using a pre-trained NLI model for more granular annotation of negative NLI instances, could enhance the performance of PI2NLI. We believe this represents a promising direction for future research.

B Training

The AdamW optimizer (Loshchilov and Hutter, 2019) is employed with a learning rate of 1e-5 and a batch size of 32. We tune the learning rate within the range of [1e-5, 2e-5, 5e-5] and choose the batch size to optimize the GPU memory utilization on a single Nvidia Tesla V100. To prevent overfitting, we adopt early stopping on the F1 score of validation for 6 epochs (Prechelt, 1998). All implementations are executed using PyTorch (Paszke et al., 2019), with pre-trained models sourced from the HuggingFace repository (Wolf et al., 2020).

In our implementation, we transitioned from a standard PI pipeline consistent with established practices in existing literature (Peng et al., 2022) to our PI2NLI. This strategic shift was executed with an emphasis on ensuring fairness and comparability across tests. Thus, our setup may even slightly favor the PI baselines. While more precise tuning of training configurations might enhance the performance of PI2NLI, our primary focus has been on validating our hypothesis. Our future work will explore optimizing these configurations to further improve performance.

C Data

The Paraphrase and Semantic Similarity in Twitter (PIT) dataset is sourced from Twitter's trending service and annotated using Amazon Mechanical Turk (Xu et al., 2015). The labels range from 0 to 5. We follow the suggested binary data processing where labels 4 and 5 indicate a paraphrase, and labels 0 through 2 do not.[6]

The Quora Question Pairs (QQP) dataset originates from the question-and-answer platform Quora, consisting of question pairs annotated for potential duplicity (Iyer et al., 2017). The dataset labels are binary, indicating whether question pairs are duplicates (TRUE) or not (FALSE).[7]

The Microsoft Research Paraphrase Corpus (MSRP) is derived from sentence pairs generated by clustering news articles using heuristic extraction and an SVM classifier, with human annotations provided (Dolan and Brockett, 2005). For this study, we adhere to the GLUE benchmark standards for processing and splitting the data (Wang et al., 2018).[8]

The PARAphrase identification based on Domain knowledgE (PARADE) dataset is tailored for PI in computer science, requiring in-depth domain knowledge (He et al., 2020). It challenges models to identify paraphrases that, despite minimal lexical and syntactic overlap, are semantically equivalent due to the specialized context of computer science. The dataset offers annotations in both four-class and binary formats, provided by annotators with domain expertise.[9] In this work, we use binary labels to maintain consistency with prior studies.

The Paraphrase Adversaries from Word Scrambling (PAWS) benchmark, including PAWS QQP and PAWS Wiki, is proposed to test models to discern semantic relationships despite superficial lexical similarities (Zhang et al., 2019). These datasets utilize word scrambling and back-translation to create adversarial examples that, while sharing high lexical overlap, differ significantly in meaning. PAWS QQP draws questions from the QQP corpus and PAWS Wiki is based on sentences from Wikipedia.[10] Labels are provided in binary format, and we follow the standard data processing protocols as originally released.[11]

[6] https://github.com/cocoxu/SemEval-PIT2015

[7] https://huggingface.co/datasets/quora

[8] https://huggingface.co/datasets/nyu-mll/glue

[9] https://github.com/heyunh2015/PARADE_dataset

[10] https://dumps.wikimedia.org/

[11] https://github.com/google-research-datasets/paws