NKT: Evidence-aware Inference for Neutralized Zero-shot Transfer

Xiaotong Feng*, **Meng-Fen Chiang***, **Wang-Chien Lee**†, **Zixin Kuang***
*{xfen974, mchi174, zkua236}@aucklanduni.ac.nz, The University of Auckland, New Zealand
†wlee@cse.psu.edu, The Pennsylvania State University, USA

Abstract

Cross-domain knowledge transfer, which has received growing research attention in natural language processing (NLP), is a promising approach for various NLP tasks such as evidence-aware inference. However, the presence of biased language in well-known benchmarks notably misleads predictive models due to the hidden false correlations in the linguistic corpus. In this paper, we propose **N**eutralized **K**nowledge **T**ransfer framework (**NKT**) to equip pre-trained language models with neutralized transferability. Specifically, we construct debiased multi-source corpora (CV and EL) for two exemplary knowledge transfer tasks: claim verification and evidence learning, respectively. To counteract biased language, we design a neutralization mechanism in the presence of label skewness. We also design a label adaptation mechanism in light of the mixed label systems in the multi-source corpora. In extensive experiments, the proposed NKT framework shows effective transferability contrarily to the disability of dominant baselines, particularly in the zero-shot cross-domain transfer setting.

Keywords: Debiased Learning, Zero-shot Transfer, Language Neutralization, Learnable Instance Re-weighting

1. Introduction

Motivations. The recent success of natural language processing (NLP) is attributed not only to advances in learning models and computational resources but also to the availability of massive, richly-labeled datasets. However, some real-world NLP tasks, due to the absence of annotated data, may rely on knowledge transfer from some richly-labeled datasets. A knowledge transfer to some unseen domain is referred to as cross-domain transfer (Feng et al., 2021), while a knowledge transfer to the seen domain is referred to as in-domain transfer. Under the settings of cross-domain transfer, the most challenging one is to transfer knowledge to a target domain without fine-tuning, which is referred to as *zero-shot* transfer.

Most prior work on NLP tasks is formulated as an isolated learning process, where the predictive model is narrowly tailored for a single task trained on a single dataset. For instance, some datasets of claim verification, which aim to determine the relationship between a claim and evidence, come without supporting evidence (Onoe et al., 2021; Sepúlveda-Torres et al., 2021) as shown in Figure 1(a). To accurately verify a claim, retrieving precise evidence that can truly support the claim is crucial. Although DPR (Karpukhin et al., 2020) supports efficient evidence retrieval under in-domain transfer (e.g., Wikipedia to Wikipedia), the learned knowledge transfer models cannot be applied to facilitate evidence retrieval for claims from different domains. This is because many richly-labeled training examples from the target domain are required for cross-domain transfer to an unseen dataset (i.e., zero-shot transfer). However, relying on human annotation to prepare labeled training data is labor-intensive. To enable *cross-domain transfer*, especially in the zero-shot setting, we propose to follow the pre-finetuning framework (Aghajanyan et al., 2021) by leveraging multi-source learning as an intermediate step between the standard pre-training and fine-tuning framework.

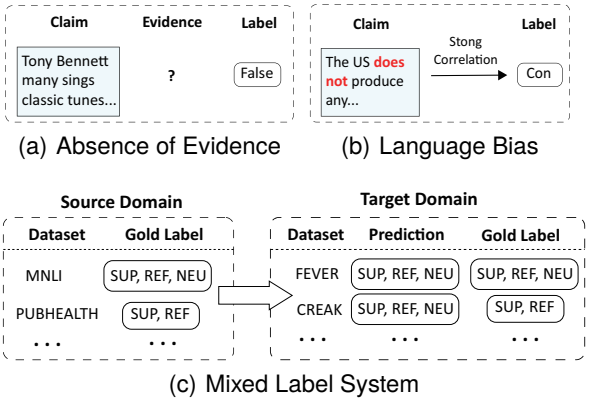


Figure 1: Motivations. (a) Supporting gold evidence is missing in the CREAK dataset, which needs automatic evidence retrieval. (b) The phrase "does not" strongly correlates with the label "CONTRADICTION" in the MNLI dataset. (c) A multi-source corpus contains two-label and three-label systems.

An issue arising in the multi-source learning for NLP is that language sources are usually biased, notably due to their varied degrees of *linguistic idiosyncrasy* (Baldwin et al., 2021; Schuster et al., 2019a; Liu et al., 2019a). The ramification of a biased language is significant when a predictive model performs incredibly well, not because of the model capability but due to the hidden false (misleading) correlations in the linguistic corpus. Figure 1(b) illustrates the false correlations between the

LREC-COLING 2024, pages 6671–6681

biased phrases and claim labels. To tackle the language bias, (Schuster et al., 2019a) formalizes a debiasing mechanism to break the false linguistic correlations with a particular label. However, the debiasing mechanism fails to deal with biased phrases under various degrees of label skewness. Therefore, a robust approach must consider label skewness when tackling hidden false correlations.

Multi-source dataset brings the complexity of *mixed label systems*, which may cause label prediction muddle. Even similar datasets may differ in label systems, e.g., a two-way task on CREAK (Onoe et al., 2021) with the label inventory {TRUE, FALSE} versus a three-way task on FEVER (Thorne et al., 2018) with the label inventory {SUP, REF, NEI}. Figure 1(c) illustrates a mixture of two label systems in the source and target domains. Prior work (Mithun et al., 2021a) usually constrains a source dataset with the same number of label types as the target task to naturally transfer developed knowledge in the pre-training phase. These approaches limit the choice of source datasets to a small extent. Other approaches (Feng et al., 2021; Wang et al., 2021) simply ignore the label adaptation problem, resulting in the disability of zero-shot transfer. To alleviate the prediction muddle, *label adaptation* is needed to acquire transferable knowledge for cross-domain knowledge transfer in multiple label systems, especially when the zero-shot transfer is demanded.

Research Objective. To enable label adaptation and neutralized knowledge transfer across different domains, we propose a **N**eutralized **K**nowledge **T**ransfer framework (**NKT**). NTK is further tailored for two exemplary neutralized knowledge transfers: *claim verification* and *evidence learning*, resulting in two variant models, **NKT+CV** and **NKT+EL**. Note that the NKT framework is task-agnostic. They can be applied to gain transferability through the pre-finetuning phase on a source corpus for diverse target tasks. Specifically, we equip a sequence-to-sequence Vanilla T5 (Raffel et al., 2020) with the neutralized knowledge transferability upon the NKT framework. Note that Vanilla T5 is designed for multi-task learning compared to other language models (Devlin et al., 2018; Liu et al., 2019b), which are designed for single-task learning. Moreover, T5 is a natural choice as a transfer learning backbone, facilitating our approach to label adaptation. To expand knowledge, we construct two multi-source corpora with richly blended benchmarks from various domains: a claim verification pre-finetuning corpus (CV), and an evidence learning pre-finetuning corpus (EL). To counteract the effect of hidden false correlations, we design a neutralization learning mechanism to adjust dependency for biased phrases as well as the instance-wise contribution toward each corpus. Afterwards, we pre-finetune Vanilla T5 on neutralized CV and EL corpus to derive respective transferable models: (i) NKT+CV with a novel label adaptation mechanism for claim verification; and (ii) NKT+EL for evidence learning. To summarize, the main contributions of this study are as follows.

- We propose a **N**eutralized **K**nowledge **T**ransfer framework (NKT) to learn neutralized knowledge from multi-source corpora, CV, and EL.
- We design a neutralization mechanism to counteract the effect of the hidden false correlations under various degrees of label skewness.
- We equip T5 Vanilla with evidence-aware inference capability by pre-finetuning NKT+CV and NKT+EL for evidence-aware claim verification and automatic evidence retrieval, respectively.
- We demonstrate the effective transferability of the NKT framework over representative baselines in extensive experiments, particularly under zero-shot cross-domain transfer settings.

2. Related Works

Claim Verification. Several studies have been conducted on claim verification to verify the authenticity of a given claim. FEVER (Thorne et al., 2018) is a widely adopted benchmark for claim verification with supporting evidence from Wikipedia. FEVEROUS (Aly et al., 2021) is constructed based on FEVER, containing not only textual sources but also tabular information. CREAK (Onoe et al., 2021) is a benchmark that requires models to have the abilities of fact retrieval and commonsense reasoning. MultiFC (Augenstein et al., 2019) is an evidence-aware multi-source benchmark collected from multiple websites such as Politifact and Snopes. COVIDFACT (Saakyan et al., 2021) and PUBHEALTH (Kotonya and Toni, 2020) are healthcare domain datasets about COVID-19.

Evidence Learning. Evidence learning performance may influence the prediction accuracy of claim verification task with fake evidence. DPR (Karpukhin et al., 2020) trains the dense representation to encode questions and their positive passages with high similarity in representations via the in-batch negative strategy. RAG (Lewis et al., 2020b) is pre-trained with a dense vector index of Wikipedia combining a sequence-to-sequence model BART (Lewis et al., 2020a) to perform the evidence retrieval and question-answering tasks.

Transfer Learning. Some efforts have been devoted to tackle the absence of labeled data. (Mithun et al., 2021b,a) focus on data and model distillation by group learning and teacher-student architecture. Instead, (Feng et al., 2021) focus on data selection with a reinforced selector to select samples from source domain similar to the target domain.

(Hardalov et al., 2021) focus on adapting labels of cross-domain stance detection with label embeddings. (Wang et al., 2021) apply meta-learning for tasks adapting and labels adapting. (Aribandi et al., 2021) pre-train a T5 model on a large-scale corpus via diverse tasks to improve the transferability.

Debiasing Methods. There exist several works on reducing the bias in a linguistic corpus. They can be divided into two classes. The first class is post-processing methods (Doherty et al., 2012; Jiang et al., 2020). After pretraining a classification model, these works consider the fairness by adjusting its output. Since the process handles correctness and fairness independently, it is hard to balance them simultaneously. Methods in the second class reduce the bias during the training process by adjusting the learning objectives (Schuster et al., 2019a; Karimi Mahabadi et al., 2020; Sanh et al., 2021). Usually, they re-weight the biased cases to reduce their influence, which may sometimes result in complicated calculations. These methods all consider the corrections between n-grams and labels. Therefore, some independent n-gram quantification methods, e.g., TD-IDF, cannot achieve the objectives (Robertson, 2004). Here we follow the method adopted for the FEVER dataset (Schuster et al., 2019a), which belongs to the second class and is model-agnostic. It re-weights the instances to adjust the new learning objective.

3. Preliminaries

3.1. Generative Language Models

The complete encoder-decoder Transformer architecture (Vaswani et al., 2017) is widely adopted to build sequence-to-sequence models (Lewis et al., 2020a; Raffel et al., 2020). Amongst them, T5 (Raffel et al., 2020) has been explored for transfer learning, where a Vanilla T5 is first pre-trained on a large-scale unlabeled dataset "Colossal Clean Crawled Corpus" (C4), with a maximum likelihood objective. It is then fine-tuned on a family of learning tasks (e.g., translation, question answering, and text classification), where each task is distinguished by a task-specific prefix as part of the input. Inspired by these successes, we adopt the Vanilla T5 as the backbone of this work. We aim to enhance its evidence-aware inference capability for claim verification and evidence retrieval via benchmark corpus in which various degrees of language bias and mixed label systems exist.

3.2. Linguistic Idiosyncrasy

To identify bias, (Schuster et al., 2019a) propose to use local mutual information (LMI), where some n-grams in a sentence appear to be falsely correlated with a particular label (e.g., REFUTE), compared with other labels (e.g., SUPPORT).

n-gram (#SUP)	LMI	n-gram (#REF)	LMI
higher longitudinal	573	has no	1,373
has higher	573	no known	1,030
can be	453	solely by	944
requires hypothalamic	420	obesity determined	944
balance requires	420	determined solely	944

Table 1: Top-5 biased n-grams ranked by LMI ($\times 10^{-6}$) on SCIFACT dataset.

$$\text{LMI}(w, l) = p(w, l) \cdot \log\left(\frac{p(l|w)}{p(l)}\right) \quad (1)$$

where $p(w, l)$ is the joint probability of the n-gram w and the label l in the corpus $\mathcal{D}$, and $p(l)$ is the probability of l in $\mathcal{D}$. Table 1 shows examples of biased n-grams ranked by LMI metric. To tackle these biases, the authors propose to lower the importance of instances containing biased n-grams in the dataset with a learnable instance weight. Consequently, the impact of biased n-grams, initially high in LMI, is reduced. Specifically, the authors estimate the impact of an n-gram (w_j) biased toward label l as follows.

$$b_{w_j}^{(l)} = \frac{\sum_{i=1}^{N} I[c_i, w_j](1 + \alpha_i) I[y_i = l]}{\sum_{i=1}^{N} I[c_i, w_j](1 + \alpha_i)}, \quad (2)$$

where α_i is a learnable instance weight associated with the claim c_i. $I[c_i, w_j]$ is an indicator function with value 1 if the claim c_i contains the n-gram w_j, and the value of the indicator function $I[y_i = l]$ is 1 if the label class y_i of the instance is l. The higher the bias $b_{w_j}^{(l)}$ of w_j toward a label l, the instance weight α_i for the claim c_i, which contains the w_j, should be lowered to decrease the numerator in Eq. (2). To further decrease the bias $b_{w_j}^{(l)}$ by increasing the denominator in Eq. 2, the weights α_i toward other labels should be increased.

To learn the best α_i for each claim instance c_i $\forall 1 \leq i \leq N$, the authors formulate an adversarial learning objective function to minimize the overall impact of notable biased n-grams as follows.

$$\min\left(\sum_{j=1}^{|V|} \max_{l}(b_{w_j}^{(l)}) + \lambda \left\|\vec{\alpha_i}\right\|_2\right), \quad (3)$$

where $|V|$ is the total number of biased n-grams which are determined based on the LMI values. The adversarial objective function searches for the maximum bias across all classes while minimizing the overall bias by adjusting α values. The L2 norm is a regularization term to penalize drastic instance weights with λ controlling their respective importance.

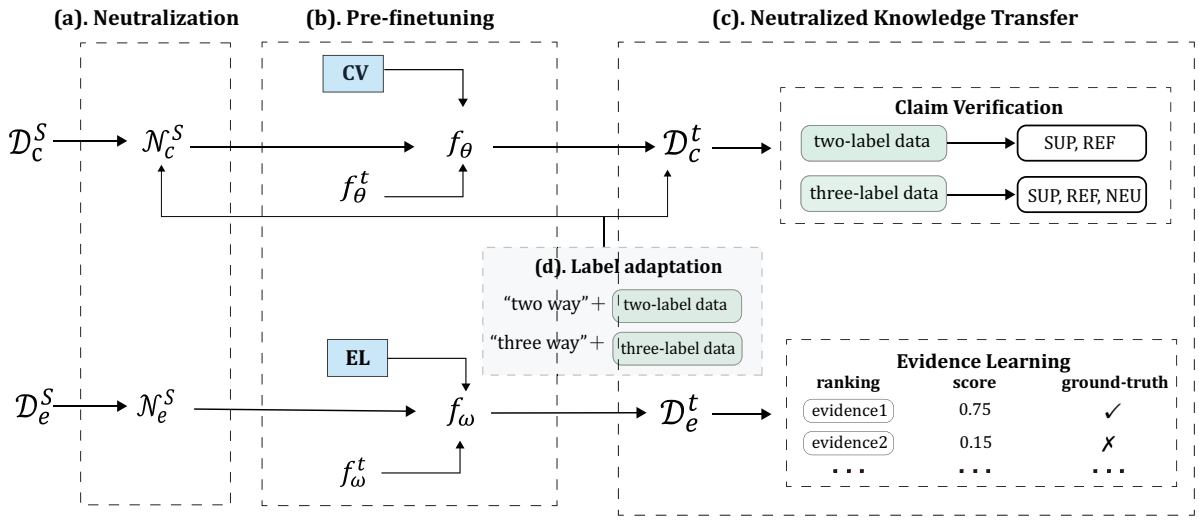


Figure 2: An overview of the NKT framework. (a) NKT neutralizes biased multi-source domains D_c^s (D_e^s) into unbiased corpora $\mathcal{N}_c^s$ ($\mathcal{N}_e^s$). (b) NKT pre-finetunes the pre-trained language model (f_θ^t,f_ω^t) on both corpora to develop NKT+CV model (f_θ) via the claim verification task and NKT+EL model (f_ω) via the evidence learning task. (c) NKT transfers neutralized knowledge with f_θ and f_ω as the initial models for target domains. NKT fine-tunes or leverages f_θ (f_ω) immediately (i.e., zero-shot) for the claim verification task in target domains with varied label systems (for automatic evidence retrieval as a ranking problem). (d) A label adaptation mechanism with proposed prefix descriptors.

After debiasing, most of the high LMI values of biased n-grams are decreased, indicating that the biased dataset has been neutralized across labels to some degree. However, Eq. (2) can not deal with n-grams under extremely unbalanced label distribution (i.e., $p(l|w_j) = 1$). To solve the problem, we propose a robust method in Section 4.1 to neutralize the biased corpus.

3.3. Zero-shot Knowledge Transfer

In-domain or cross-domain transfer can be achieved to some extent by a two-stage learning process, where the pre-training stage learns parameterized knowledge θ^s from the source domain corpus, and the fine-tuning stage refines θ^s with the target domain corpus.

We hypothesize that the transferability of neutralized knowledge can be achieved via a *pre-finetuning* phase, an intermediate step between the pre-training and fine-tuning stages. In particular, the cross-domain transfer can be further distinguished into fine-tuning and *zero-shot transfer* settings, depending on whether the final model learns from any training samples in the target corpus. The zero-shot transferability is harder yet vital, especially when computational resources are too limited to afford entire fine-tuning stage in the target domain.

3.4. Problem Statement

Given a source corpus $\mathcal{D}^s$, the goal is to enhance the transferability of neutralized knowledge for in-domain and cross-domain transfer in the target corpus $\mathcal{D}^t$ in both fine-tuning and zero-shot settings.

4. Proposed Model

In this paper, we present the **N**eutralized **K**nowledge **T**ransfer framework (NKT) built upon T5 for neutralized knowledge transfer on two evidence-aware inference tasks: *claim verification* and *evidence learning*. The NKT framework consists of four modules: a) *Neutralization*, b) *Pre-finetuning*, c) *Neutralized Knowledge Transfer*, and d) *Label adaptation*. The neutralization module debiases the constructed corpora for inference tasks. The pre-finetuning module learns evidence-aware claim verification knowledge (evidence learning knowledge) upon pre-trained T5 Vanilla on neutralized claim verification source corpus (neutralized evidence learning source corpus). The neutralized knowledge transfer module transfers the learned models for evidence learning and claim verification. The label adaptation module enables claim verification for mixed label systems in the target domain. Figure 2 gives an overview of the NKT framework.

4.1. Neutralization Learning

Motivation. As discussed earlier, the debias approach (Schuster et al., 2019b) performs poorly when the label distribution is highly unbalanced. In other words, when w_j biases toward class l with a conditional probability of 1 (i.e., w_j never appears in other classes), Eq. (2) fails to reduce the impact of biased w_j. Consequently, it also fails to minimize the importance of those instances containing w_j.

n-gram Bias Estimation. To amend the weakness, we introduce a *global bias* to reinforce weight adjustment for instances containing the biased n-gram w_j even if the n-gram has a conditional probability 1 toward class l as follows.

$$g_{w_j}^{(l)} = \frac{\sum_{i=1}^{N} I[c_i, w_j](1+\alpha_i) I[y_i = l]}{\sum_{i=1}^{N}(1+\alpha_i)}, \quad (4)$$

where the global bias of w_j is the conditional probability toward class l normalized over the summation of all instance weights in the corpus. When the global bias of w_j is encouraged to decrease, the weight α_i of the instance containing the biased n-gram w_j is encouraged to reduce the numerator in Eq. (4). To further decrease the global bias by increasing the denominator in Eq. (4), the weights α of all instances are encouraged to be increased. We then define the overall bias of w_j by considering the global bias to tackle arbitrary label skewness as follows.

$$bias_{w_j}^{(l)} = b_{w_j}^{(l)} + (b_{w_j}^{(l)} - \frac{1}{|L|}) \times g_{w_j}^{(l)}, \quad (5)$$

where $|L|$ is the number of classes in the corpus. If the estimated bias is unbalanced (i.e., $b_{w_j}^{(l)} \in (\frac{1}{|L|}, 1]$), the importance toward class l for instances containing w_j is encouraged to decrease. Complementarily, the importance toward other classes for instances containing w_j is encouraged to increase before increasing the importance of all instances. If the estimated bias is balanced (i.e., $b_{w_j}^{(l)} = \frac{1}{|L|}$), the global bias makes no difference.

Instance Re-weighting. Given the amended $bias$ estimation, we calibrate the adversarial learning objective function in Eq. (3) to search for the optimal weight α_i for each instance c_i in the corpus as follows.

$$\min \left(\sum_{j=1}^{|V|} \max_{l}(bias_j^{(l)}) + \lambda \left\| \vec{\alpha_i} \right\|_2 \right), \alpha_i \in (-1, 1], \quad (6)$$

where the overall bias defined by the set of the maximum $bias$ estimations across all classes is minimized. Note that the regularized instance weight $\alpha_i \in (-1, 1]$ differs from $\alpha_i \in [0, \infty]$ in Eq. (3). λ is to control the degree of neutralization when $p(l|w_j) = 1$ by penalizing the value of α in the global bias $g_{w_j}^{(l)}$. Finally, we obtain a neutralized corpus, denoted as $\mathcal{N}^s$, for a given source corpus $\mathcal{D}^s$.

4.2. Claim Verification Transfer

Given a claim and the associated evidence, we aim to equip our model with claim verification capability and transfer the capability to the target domain. For this purpose, we first enrich the knowledge of our source domain corpus with both general knowledge and specific expertise and then develop claim verification capability for the target domains via pre-finetuning framework.

CV Corpus Construction. Specifically, we integrate representative datasets into a benchmark corpus (CV) as the source domain, including general knowledge (FEVER (Thorne et al., 2018) and MNLI (Williams et al., 2018)) and specific expertise in science and healthcare domains (SCIFACT (Wadden et al., 2020) and PUBHEALTH (Kotonya and Toni, 2020)). The CV source corpus is prepared to learn claim verification knowledge.

Pre-finetuning to Transfer. Given the neutralized source corpus $\mathcal{N}_c^s = \{(c_i, e_i, l_i, \alpha_i) | \forall 1 \leq i \leq |\mathcal{N}_c^s|\}$, where each sample consists of a claim c_i, an evidence e_i associated with the claim c_i, a class label $l_i \in \{L_1^s, L_2^s, ...\}$ and an instance weight α_i, obtained via neutralization learning on CV corpus. Our goal is to pre-finetune a parameterized model f_θ on $\mathcal{N}_c^s$ to accurately predict the claim label l_i for a given claim c_i. An example of a target task is to leverage f_θ as the initial model to further fine-tune or straightforwardly perform on the target corpus $\mathcal{D}_c^t$ (zero-shot) to determine the claim label $l_j \in \{L_1^t, L_2^t, ...\}$ for a given claim $c_j \in \mathcal{D}_c^t$.

Label Adaptation Prefix. A model trained on a source corpus with an m-way label system, comprising in total m label classes, can not directly transfer to a target dataset with an n-way label system containing totally n label classes. To address the issue, we modify the T5 model by introducing a new prefix descriptor to enable cross-domain label adaptation. Specifically, we introduce an additional label adaptation prefix to the original input sentence before forwarding the data for the pre-finetuning stage. For each dataset in our corpus, a label adaptation prefix is provided as part of the input to indicate the label system the dataset belongs to. The input format for "m-way" instances, indicating the dataset has m unique label classes, is described as follows.

```
m-way: [claim] + [evidence]
```

For example, the prefix "three-way" is added to each instance in FEVER, indicating it is a three-label dataset. As such, the label adaptation prefix informs the NKT+CV model, which label system to adapt to during the pre-finetuning and inference stages. Note that for those cases where the label classes of the target dataset are more than that of the source dataset, we use the source label classes to ensure complete transfer.

4.3. Evidence Learning Transfer

To address the issue of missing evidence annotations, we aim to equip our model with evidence retrieval capability to supplement claims with gold evidence in target domains. This is referred to as the transferability of evidence learning.

EL Corpus Construction. Specifically, we construct a multi-source corpus for evidence learning (EL), including the same sources as the CV (i.e., FEVER, MNLI, PUBHEALTH, and SCIFACT). To enhance the learning process, we also augment EL with negative samples.

Negative Sample Augmentation. Given a claim and the associated evidence in EL, we first collect samples with labels in {SUP, REF}. Then, we randomly selected ten pieces of evidence from other claims for each claim to pair with the claim as negative samples. After the augmentation process, the number of negative samples is exactly ten times that of positive samples.

Pre-finetuning to Transfer. Given the neutralized EL source corpus $\mathcal{N}_e^s = \{(c_i, p_i, l_i, \alpha_i)|\forall 1 \leq i \leq |\mathcal{N}_e^s|\}$, where each sample consists of a claim c_i, an explainable passage p_i, a gold entailment label $l_i \in \{\text{POS}, \text{NEG}\}$ indicating if p_i truthfully (l_i=POS) or falsely (l_i=NEG) entails the claim c_i, and an importance weight α_i, obtained via neutralization learning on EL corpus. Our goal is to pre-finetune a parameterized model f_ω on $\mathcal{N}_e^s$ to accurately predict the entailment label l_i for the claim-evidence pair (c_i,p_i). An example of a target task is to automatically retrieve evidence for a given claim in the target corpus $\mathcal{D}_e^t$. This can be cast as an evidence ranking problem for a claim $c_j \in \mathcal{D}_e^t$ by leveraging f_ω as the initial model for different transfer settings.

4.4. Re-weighted Learning Objective

To factor in the instance importance (α_i), we define a re-weighted cross entropy loss to pre-finetune NTK+CV and NTK+EL as follows.

$$\mathcal{L}_w = -\sum_{i=1}^{N}(1+\alpha_i)\sum_{l=1}^{|L|} y_i^{(l)} \log(p_i^{(l)}) \qquad (7)$$

where $\mathcal{L}(c_i, y_i)$ is the cross entropy loss, $p_i^{(l)}$ is the softmax probability for the l-th class, and $y_i^{(l)}$ is the actual class label. NKT+CV (f_θ) is optimized via class verification task by minimizing $\mathcal{L}_w$ on $\mathcal{N}_c^s$ = $\{(c_i, e_i, y_i^{(l)}, \alpha_i)\}$, where $y_i^{(l)} \in \{L_1^s, L_2^s, ...\}$. NKT+EL ($f_\omega$) is optimized via evidence learning task by minimizing $\mathcal{L}_w$ on $\mathcal{N}_e^s$ = $\{(c_i, p_i, y_i^{(l)}, \alpha_i)\}$, where $y_i^{(l)} \in$ {POS,NEG}.

5. Experiments

In this section, we conduct experiments to qualitatively and quantitatively study the effectiveness of NKT via claim verification and evidence learning tasks. The knowledge transferability (RQ1) verifies the transferability of NKT. The ablution study (RQ2) addresses the contribution of label adaptation and neutralization learning without instance reweighting. The neutralization study (RQ3) explores whether words with bias are responsibly adjusted by the proposed neutralization module at both the word and instance levels.

CV	Domain	#SUP	#REF	#NEI
FEVER	Wikipedia	80,035	29,775	35,639
MNLI	Inference	130,899	130,903	130,900
PUBHEALTH	Healthcare	5,078	3,001	0
SCIFACT	Science	832	463	0
EL	Domain	#POS	#NEG	
FEVER	Wikipedia	90,528	905,280	
MNLI	Inference	261,802	2,618,020	
PUBHEALTH	Healthcare	8,079	80,790	
SCIFACT	Science	1,295	12,950	

Table 2: Data statistics of the source corpora.

Transfer Settings		In-Domain	Cross-Domain (Zero-shot)	
Target Corpus		FEV	CREAK	COVID-FACT
Few-shot	KGAT	.8033	-	-
	GEAR	.7782	-	-
	BERT	.8543	.496	.6782
	RoBERTa	.5	.496	.6782
	T5 (Vanilla)	.8591	.7162	.6806
	NKT+CV + L	**.8663**	**.7987**	**.7946**
Zero-shot	BERT	-	-	-
	RoBERTa	-	-	-
	T5 (Vanilla)	-	-	-
	NKT+CV + L	**.891**	**.7549**	**.5099**

Table 3: Claim verification task.

5.1. Datasets

Claim Verification. For claim verification pre-finetuning corpus (CV), we collect the training set of FEVER, MNLI, PUBHEALTH, and SCIFACT. For label alignment, we collect samples from PUBHEALTH with labels in { FALSE, TRUE}, and samples from SCIFACT (Wadden et al., 2020) with labels in {SUP, REF}. To evaluate the claim verification transferability, we collect (i) a validation set of FEVER for the in-domain transfer settings and (ii) a validation set of the CREAK and a test set of COVID-FACT for cross-domain transfer settings.

Evidence Learning. For evidence learning pre-finetuning corpus (EL), we collect the training sets of FEVER, MNLI, PUBHEALTH, and COVID-FACT. To evaluate the evidence learning transferability, we collect (i) validation sets from MNLI-matched, FEVER, and test sets from PUBHEALTH for in-domain transfer settings and (ii) validation sets from MNLI-mismatched and test sets from COVID-FACT for cross-domain transfer setting. Note that the MNLI-mismatched validation set is considered a cross-domain transfer setting, as some genres do not appear in the MNLI training set. Table 2 summarizes the detailed statistics of CV and EL corpora.

Transfer Settings	In-Domain									Cross-Domain (Zero-shot)					
Target Corpus	MNLI-matched			FEVER			PUBH			MNLI-mismatched			COVID-FACT		
Metric	P@1	R@1	F1@1	P@1	R@1	F1@1	P@1	R@1	F1@1	P@1	R@1	F1@1	P@1	R@1	F1@1
Random	.002	.002	.002	.0	.0	.0	.001	.001	.001	.0	.0	.0	.0074	.0074	.0074
DPR	.375	.375	.375	.553	.553	.553	.3171	.3171	.3171	.357	.357	.357	.1733	.1733	.1733
BERT + EL	.928	.928	.928	.843	.843	.843	.8267	.8267	.8267	.9	.9	.9	.7351	.7351	.7351
RoBERTa + EL	.921	.921	.921	.821	.821	.821	**.8379**	**.8379**	**.8379**	.908	.908	.908	.6733	.6733	.6733
T5 (Vanilla)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
NKT+EL - N	**.928**	**.928**	**.928**	.841	.841	.841	.8349	.8349	.8349	**.918**	**.918**	**.918**	.7871	.7871	.7871
NKT+EL	.914	.914	.914	**.885**	**.885**	**.885**	.8328	.8328	.8328	.904	.904	.904	**.8119**	**.8119**	**.8119**

Table 4: Evidence learning task with ablation on neutralization learning without instance re-weighting (-N).

5.2. Experimental Settings

Baselines. To evaluate NKT+CV, we compare with (i) transformer-based models (Vanilla T5, BERT, and RoBERTa), and (ii) graph-based models on evidence-aware claim verification (KGAT, GEAR) on FEVER dataset. To evaluate NKT+EL, we compare with DPR and a random retriever (Random) by randomly selecting top-k sentences from the corpus.To ensure a fair comparison with all baselines, original test sets of the datasets are adopted, which helps to examine the model's ability to withstand biased language within the corpus.

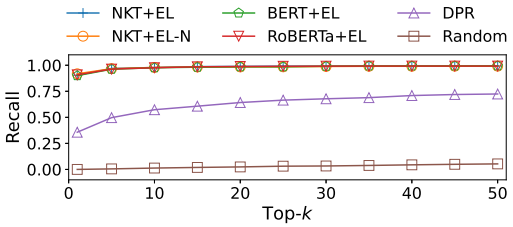
Metrics. For claim verification evaluation, we adopt the classification accuracy. For evidence learning evaluation, we adopt precision (P@k), recall (R@k), and F1-score (F1@k), where k is the number of retrieved items. Note that P@1 is the proportion of relevant recommended items in the top-1 set. If P@1=1/1, the evidence is considered valid. Conversely, if P@1=0/1, it implies that the evidence is incorrect. This relationship also applies to recall, given that each claim in our dataset corresponds to one piece of evidence, resulting in a common denominator of 1 for both recall and precision.

5.3. Knowledge Transferability (RQ1)

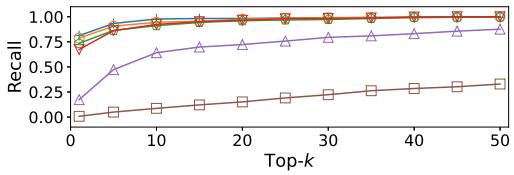
Claim Verification. As shown in Table 3, NKT+CV+L (with label adaptation) outperforms all baselines in both few-shot and zero-shot settings: (i) T5 Vanilla, BERT, and RoBERTa on all target datasets; and (ii) KGAT and GEAR on FEVER with fine-tuning. The NA in Table 3 indicates the inability of BERT, RoBERTa, and T5 Vanilla to perform transfer learning directly.

Evidence Learning. We conduct experiments to verify if NKT+EL enhances the transferability for evidence learning. Table 4 shows that our NKT+EL model performs best across the target corpus against Random and DPR. Besides, we find that evidence learning transferability works well for transformer-based models. In the in-domain transfer setting, NKT+EL achieves 0.8328 in F1@1, whereas DPR only achieves 0.3171 in F1@1 on PUBHEALTH. In the cross-domain transfer setting, NKT+EL achieves 0.8119 in F1@1, while DPR only achieves 0.1733 in F1@1 on COVID-FACT.

Figure 3 gives a closer look at cross-domain performance, where NKT+EL consistently outperforms the DPR with a significant margin across varying k on MNLI-mismatched and COVID-FACT. For instance, when $5 \leq k$, NKT+EL achieves over 0.9



(a) MNLI-mismatched



(b) COVID-FACT

Figure 3: Cross-domain evidence learning task.

Transfer Settings		In-Domain	Cross-Domain (Zero-shot)	
Target Corpus		FEV	CREAK	COVID-FACT
Few-shot	NKT+CV - N	**.8762**	.8118	.7847
	NKT+CV	.8686	.7834	.7946
	NKT+CV + L - N	.881	**.822**	**.807**
	NKT+CV + L	.8663	.7987	.7946
Zero-shot	NKT+CV - N	**.8947**	.5711	.4802
	NKT+CV	.8915	.5828	.453
	NKT+CV + L - N	.8765	.7338	**.547**
	NKT+CV + L	.891	**.7549**	.5099

Table 5: Claim verification task with ablations on label adaptation (+L) and neutralization learning without instance re-weighting (-N).

CV: Claim with #REF	α	**CV**: Claim with #SUP	α
Molyvos is a run down town in the center of the region, not popular at all with tourists.	-0.813	They were not pleased that Russia spear-headed the peace talks at all.	0.945
State size does not affect the rate of unfiled returns.	-0.592	The Dept. of Defense does not have the criteria, ...	0.832
EL: Evidence with #POS	α	**EL**: Evidence with #NEG	α
Increased risk of breast cancer was noted with increased birthweight (relative risk [RR] 1.15 [95% CI 1.09-1.21]), ...	-0.956	Protective hazard ratios (HRs) were detected for bariatric surgery for incident T2DM, 0.68 (95% CI 0.55-0.83), ...	0.965
The relative risk estimate , ... for breast cancer was 1.23 (95% confidence interval 1.13-1.34).	-0.838	Nine studies investigated the association between, ... 95% confidence interval 1.31 to 2.29).	1

Table 6: Re-weighting for claim verification and evidence learning tasks after neutralization learning.

in Recall, while DPR only reaches $\approx$ 0.7 in Recall at $k = 50$ on MNLI-mismatched. Note that T5 Vanilla pre-trained on unlabeled text corpus can not directly perform the entailment ranking task. Contrarily, pre-finetuning on EL enables NKT+EL to direct transfer to the target entailment ranking task with an outstanding performance against DPR. Besides, we pre-finetune BERT and RoBERTa on EL, which results in BERT+EL and RoBERTa+EL models, respectively. Their performance is close to NKT+EL on some target corpus, suggesting the effectiveness of the pre-finetuning framework in enhancing versatile transferability.

5.4. Ablation Study (RQ2)

Claim Verification. In the zero-shot cross-domain transfer setting, Table 5 shows that the accuracy of NKT+CV on COVID-FACT has decreased to a large extent due to the exclusion of neutralization. In contrast, the accuracy of NKT+CV on CREAK increases by 0.2. In fine-tuning settings, excluding neutralization brings little gains in accuracy (i.e., 0.2 increase in CREAK and 0.1 increase in COVID-FACT). The reason why the neutralization mechanism can not always have a positive impact on a target corpus may be that the target dataset is also biased. For example, the bi-gram "can be" biases toward "SUP" in MNLI, and the bi-gram "does not" biases toward "REF" in MNLI and FEVER. Contrarily, COVID-FACT is not biased, leading to improved prediction accuracy in a zero-shot setting. In Table 5, we notice that neutralization (N) may not consistently improve classification accuracy. This is attributed to the nature of neutralization, which involves breaking down spurious correlations to make predictions based on semantic relevance rather than relying on biased language cues.

In the zero-shot cross-domain transfer setting, Table 3 shows that the accuracy for both CREAK and COVID-FACT has marginally decreased due to the exclusion of label adaptation. According to our analysis, $\approx$ 2.5% predictions (NKT+CV) on COVID-FACT and $\approx$ 21.88% predictions (NKT+CV) on CREAK are not in respective label systems. All predictions (NKT+CV+L) on each benchmark are in the label systems, suggesting the effectiveness of our label adaption mechanism.

Evidence learning. Table 4 shows that excluding the neutralization mechanism causes $\approx$ 0.2 decline in Recall on COVID-FACT.

5.5. Debiased Learning (RQ3)

Claim Verification. Table 6 gives examples of re-weighted instances with the biased bi-grams highlighted in blue, e.g., "at all" and "does not" are biased toward "REF" in MNLI. Thus, the instances containing any of them with "REF" receive lower weights. Contrarily, the weights for instances containing any of them with the label "SUP" are increased.

Evidence Learning. Likewise, "95 CI" and "95 confidence" are biased toward "POS" in SCIFACT. Thus, instances containing any labeled with "POS" receive lower weights ≈ -1, while their weights toward "NEG" increase ≈ 1.

6. Conclusion

We propose a novel NKT framework to tackle the scarcely labeled data and the biased data, particularly in zero-shot cross-domain transfer setting. We equip a Vanilla T5 with neutralized knowledge via robust neutralization mechanism by pre-finetuning NKT+CV (NKT+EL) on neutralized CV (EL) corpus with proposed label adaptation prefix. The results show that the NKT framework effectively enhances the transferability for both tasks.

7. Limitations

The inefficiency of NKT+EL model needs attention despite its effectiveness. Diversifying the target domains is another agenda as we primarily focus on scientific domains. A robust neutralization is a must if adversarial samples exist.

8. Bibliographical References

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. *arXiv preprint arXiv:2101.11038*.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: fact extraction and verification over unstructured and structured information. In *NIPS*.

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Prakash Gupta, Kai Hui, Sebastian Ruder, and Donald Metzler. 2021. Ext5: Towards extreme multi-task scaling for transfer learning. *CoRR*, abs/2111.10952.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. 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)*.

Timothy Baldwin, William Croft, Joakim Nivre, and Agata Savary. 2021. Universals of linguistic idiosyncrasy in multilingual computational linguistics (dagstuhl seminar 21351). *Dagstuhl Reports*, 11(7):89–138.

Jifan Chen, Shih-ting Lin, and Greg Durrett. 2019. Multi-hop question answering via reasoning chains. *arXiv preprint arXiv:1910.02610*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*.

Neil A. Doherty, Anastasia V. Kartasheva, and Richard D. Phillips. 2012. Information effect of entry into credit ratings market: The case of insurers' ratings. *Journal of Financial Economics*, pages 308–330.

Lingyun Feng, Minghui Qiu, Yaliang Li, Haitao Zheng, and Ying Shen. 2021. Wasserstein selective transfer learning for cross-domain text mining. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. *CoRR*, abs/2104.07467.

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. 2020. Wasserstein fair classification. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 862–872.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *CoRR*, abs/2004.04906.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. *CoRR*, abs/2010.09926.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401.

Shijia Liu, Hongyuan Mei, Adina Williams, and Ryan Cotterell. 2019a. On the idiosyncrasies of the Mandarin Chinese classifier system. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (NAACL-HLT)*, pages 4100–4106.

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

Mitch Paul Mithun, Sandeep Suntwal, and Mihai Surdeanu. 2021a. Data and model distillation as a solution for domain-transferable fact verification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Mitch Paul Mithun, Sandeep Suntwal, and Mihai Surdeanu. 2021b. Students who study together learn better: On the importance of collective knowledge distillation for domain transfer in fact verification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6968–6973, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yasumasa Onoe, Michael Zhang, Eunsol Choi, and Greg Durrett. 2021. Creak: A dataset for commonsense reasoning over entity knowledge. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

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(140):1–67.

Stephen E. Robertson. 2004. Understanding inverse document frequency: on theoretical arguments for idf. *J. Documentation*, 60:503–520.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*.

Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M Rush. 2021. Learning from others' mistakes: Avoiding dataset biases without modeling them. In *International Conference on Learning Representations*.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin c! robust fact verification with contrastive evidence. *CoRR*, abs/2103.08541.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019a. Towards debiasing fact verification models. In *Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Tal Schuster, Darsh J. Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay. 2019b. Towards debiasing fact verification models. *CoRR*, abs/1908.05267.

Robiert Sepúlveda-Torres, Marta Vicente, Estela Saquete, Elena Lloret, and Manuel Palomar. 2021. Exploring summarization to enhance headline stance detection. In *International Conference on Applications of Natural Language to Information Systems*, pages 243–254. Springer.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 ((NAACL-HLT))*.

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 *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Guojia Wan, Shirui Pan, Chen Gong, Chuan Zhou, and Gholamreza Haffari. 2020. Reasoning like human: Hierarchical reinforcement learning for knowledge graph reasoning. In *IJCAI*. International Joint Conferences on Artificial Intelligence Organization.

Chengyu Wang, Haojie Pan, Minghui Qiu, Jun Huang, Fei Yang, and Yin Zhang. 2021. Meta distant transfer learning for pre-trained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9742–9752, Online and Punta Cana, Dominican Republic.

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 (NAACL-HLT).