Contrapositive Local Class Inference

Omid Kashefi and **Rebecca Hwa**
School of Computing and information
University of Pittsburgh
{kashefi, hwa}@cs.pitt.edu

Abstract

Certain types of classification problems may be performed at multiple levels of granularity; for example, we might want to know the sentiment polarity of a document or a sentence, or a phrase. Often, the prediction at a greater-context (e.g., sentences or paragraphs) may be informative for a more localized prediction at a smaller semantic unit (e.g., words or phrases). However, directly inferring the most salient local features from the global prediction may overlook the semantics of this relationship. This work argues that inference along the contraposition relationship of the local prediction and the corresponding global prediction makes an inference framework that is more accurate and robust to noise. We show how this contraposition framework can be implemented as a transfer function that rewrites a greater-context from one class to another and demonstrate how an appropriate transfer function can be trained from a noisy user-generated corpus. The experimental results validate our insight that the proposed contrapositive framework outperforms the alternative approaches on resource-constrained problem domains.[1]

1 Introduction

Many NLP applications analyze a piece of text at multiple levels. For example, a product review might be analyzed for whether it is informative overall; its paragraphs might be analyzed for whether they are relevant to certain aspect of the product; or its words might be analyzed for whether they express intense emotions. These classification tasks of varying scopes of context are often related. For instance, one might posit that a review containing many words carrying extreme emotions might not be very informative. In some cases, the same class prediction task (e.g., sentiment polarity) might be asked at both the global and local levels.

Classification at the more global scope depends on clues gleaned at the local level; however, classification at the local scope is often more difficult because the roles of the words and phrases vary depending on the broader context in which they are used. Therefore, a straightforward approach, such as lexicon lookup, which require domain-specific dictionaries, may not correctly identify the intended usage. While direct supervision may better take the context into account, it relies on the availability of the training corpus containing class labels for each local scope. This poses a significant bottleneck for new applications for which these resources are not widely available.

To address this problem, we exploit the relationship between the global and local classification problems and propose a training framework to infer local predictions from the corresponding global label, which may be easier to obtain. For example, suppose someone reported a social media comment as inappropriate (*global classification*), if we find out which word(s) in the comment are contributed the most to the user's decision to report the comment, we already found the inappropriate words (*local classification*) and confirms whether the sentence is correctly reported or not.

In prior work, researchers observed that local predictions might serve as *rationales* (Zaidan et al., 2007; Lei et al., 2016; Bao et al., 2018) for the global problem. For example, if a review is classified as positive, there must be some words within that review that are also positive such that they serve as the rationale for the overall prediction. Thus, if rationales can be identified for the global classification task, their local labels would be inferred to be the same as the global label. We refer to this mode of local prediction as **direct inference**. However, prior work also suggested that even modern global classifiers are still sensitive to noise, so such direct inference may fail to identify the most semantically relevant rationales. For ex-

[1] The code is available at: https://github.com/omidkashefi/contrapositive-inference

ample, irrelevant words or punctuation marks are more influential to the decision of neural text classifiers than verbs or other semantically related textual units (Mudrakarta et al., 2018); or the presence of snow in an image is the main feature to distinguish huskies from wolves rather than the features related to the animal themselves (Ribeiro et al., 2016).

In this work, we argue that this semantic relationship can be made more robust by enforcing the **contrapositive** constraint between the local prediction and its corresponding global prediction. That is, suppose we know that an instance is globally predicted to belong to some $Class\ A$ and that some local portion l contributed the most to that prediction (the rationale); if l is replaced so as to *negate* its semantic contribution, then the global label should also change (e.g. now belongs to some $Class\ B$). We propose that, *if, and only if*, the contrapositive constraint is satisfied should the local prediction be inferred from the global problem.

Modeled after style transfer and controlled text generation methods, we propose to implement the contrapositive inference scheme as a *transfer function* for global class transference. We first cast the problem as a rewriting exercise: rewrite the original global instance, which belong to some $Class\ A$, so that it becomes more likely to belong to $Class\ B$; then, those local textual parts that were changed in the rewriting are likely to be the heavy contributor to $Class\ A$ so we may infer the same class prediction for them. To train an appropriate transfer function, however, we need a corpus of training examples for the global prediction. For some problem domains, these resources are already available. For low-resource problem domains, where such annotated corpora are not available, we show how domain-inspired textual data augmentation can facilitate the training of the transfer function.

The proposed prediction framework is evaluated on several problem domains: sentiment analysis (as a resource-rich problem domain), and semantic pleonasm detection and specificity detection (as two resource-constrained domains). Results validate our insights about inferring local class labels from their corresponding global prediction labels and our proposed approach significantly outperform the alternative methods. We also demonstrate the robustness of the contrapositive local prediction to the noisy data and show that an appropriate transfer functions can be trained from corpora generated from heuristic-driven data augmentation schemes.

2 Inference by Contraposition

For many local prediction tasks, there is a corresponding, often easier to learn, global prediction task. Global prediction is relatively easier, in part, because it is attempting to classify a greater-context, which is semantically more distinctive than a localized smaller-context. Moreover, when a classification task could be performed at multiple levels of granularity (e.g., paragraphs and sentences), there is a much higher chance of having training corpora with label annotation at a larger text span than a smaller text span, which makes the more global version of the task more feasible.

The semantic purport of a greater-context is deriving from the co-occurrence of smaller semantic units. This implies that there is a semantic relation between the local and global predictions that could be used to infer the harder-to-learn local prediction from the corresponding easier-to-learn global prediction. It must be noted that prediction inference from the global prediction is only feasible for significantly contributing local features. We define the **direct inference** of local prediction from the global prediction as follows:

> $Global \rightarrow Local$: if the global prediction for a greater-context be some $Class\ A$, there exists a smaller local portion that significantly influenced the global prediction in the first place so the local prediction for it could infer the same class label as the global prediction.

However, it is reported that the classifiers might sometimes fail to learn the most semantically related features and make predictions based on just salient ones (Ribeiro et al., 2016; Mudrakarta et al., 2018; Jain and Wallace, 2019). We believe adding an extra constraint to the inference, while keeping the relationship between the local and corresponding global prediction intact, can improve the identification of the semantically related local features, therefore, we introduce the **contrapositive inference** of local prediction from the global prediction as follows:

> $\neg Local \rightarrow \neg Global$: the prediction for local portion of a context could infer the same class label as the global prediction, if, and only if, negating the semantic contribution of that smaller portion, negates the global prediction.

In the next section, we describe how this inference scheme could be applied to the discrete predictions in NLP problems.

2.1 Adaptation to NLP

The local segments of a text could serve as the features in the training process of the global prediction task, as given in Equation 1, where $\mathcal{F}$ is the global prediction function, y_i is the corresponding class label of the greater-context C_i in the training corpus, l_{ij} is the local features within the greater-context C_i, ${l_{ij}}^v$ is some vector representation of it and w_{ij} is its weight, $\mathcal{L}$ is some loss function, and b is the bias value.

$$\begin{aligned} L_C &= \min_{\mathcal{F}} \frac{1}{N} \sum_i \mathcal{L}(y_i, \mathcal{F}(C_i)) \\ &= \min_{\mathcal{F}} \frac{1}{N} \sum_i \mathcal{L}(y_i, \frac{1}{|C_i|} \sum_j w_{ij} {l_{ij}}^v + b) \end{aligned} \tag{1}$$

Now, if we can find the local feature l_{ij} that is mainly responsible for making the greater-context C_i belong to the class y_i, we may infer the same class label for that local feature as well. The *direct inference* involves finding the most contributing local features directly from the global prediction function. A straightforward way to implement this scheme is to take the local feature with the *highest weight* as the most contributing one as follows:

$$\begin{gathered} \textit{Direct Inference: Examine Weights} \\ \exists\, l_{ij} \mid w_{ij} = \arg\max_j \overrightarrow{w_i} \end{gathered} \tag{2}$$

Another popular approach to implement direct inference is to consider the contributing local features as the *rationales* for the global prediction. Rationales are defined as the reason behind the label annotation for the global prediction (Zaidan et al., 2007) and mostly used to improve the classification of the greater-context (Marshall et al., 2016; Zhang et al., 2016; Strout et al., 2019; Du et al., 2019). However, some studies have tried to develop systems for automatic extraction of the rationales (Lei et al., 2016; Ehsan et al., 2018) as the smaller text span that could replace the whole greater-context, while keeping the global-prediction intact:

$$\begin{gathered} \textit{Direct Inference: Rationale} \\ \exists\, l_{ij} \mid \mathcal{F}(l_{ij}) \approx \mathcal{F}(C_i) \end{gathered} \tag{3}$$

Alternatively, the contrapositive inference involves finding an adversarial alternative with a negated semantic contribution for local features and reevaluating the global prediction for the resulting greater-context as in Equation 4, where ${l_{ij}}^*$ is the adversarial alternative of the l_{ij}, C_i^{j*} is the corresponding greater-context when l_{ij} is replaced with ${l_{ij}}^*$. The negated global prediction, could be approximated as $\neg y_i \approx 1 - y_i$ for binary classification problems.

$$\begin{gathered} \textit{Contrapositive Inference} \\ \exists\, l_{ij},\ \exists {l_{ij}}^* \mid \mathcal{F}({C_i}^{j*}) = \neg y_i \end{gathered} \tag{4}$$

Implementing contrapositive inference, however, is not as straightforward as direct inference. A bottom-up approach requires calculating an adversarial semantic alternative for each local feature to assess their contribution, which might be very complicated and resource-intensive. Instead of that, we adapt a simpler top-down approach that can be seen as a generalization of machine translation, or as a form of *style transfer* (Bowman et al., 2016; Shen et al., 2017; Yang et al., 2018; Prabhumoye et al., 2018; Zhang et al., 2019).

We aim to develop a *transfer function* (see Section 3) that rewrites an arbitrary greater-context (e.g., a sentence) known to be in one class (e.g., $\mathit{Class\ A}$) into a corresponding text in another class (e.g., $\mathit{Class\ B}$); the smaller local parts (e.g., words or phrases) that changed during this global rewriting and class transference process are deemed to be the ones that contribute the most to the greater-context's class label, so may infer the same class label as well.

As an illustrative example, consider the sentiment classification problem. The transfer function might rewrite a *positive* sentence: "the food was great" into a *negative* sentence "The food was awful." While the overall sentence length is the same, the learned transfer function chose to replace "great" with "awful," therefore "great" is likely to be a sentiment expressing word. By casting the problem as a style transfer task, we avoid the thorny tasks of quantifying fluency and meaning retention and even if the transfer function does not provide a correct semantic negation, for example, rewrites the "great" as "cold" instead of "awful", we are not concerned with whether "The food was cold" is a meaningful or even fluent sentence; we only care that this new sentence is now classified as *negative*, so it reveals something about the word "great" in the old sentence. This property may also improve the robustness of our approach to some level of noise in the data.

Our proposed contrapositive inference approach is still a complicated method. By relaxing the global prediction negation requirement of the contraceptive inference into maximum prediction deviation, a lighter and easier-to-implement version of it could be seen as a *leave one out (LOO)* baseline. That is, if a local feature be a significant contributor to the global prediction, then removing it from the greater-context should cause a larger prediction deviation from the original prediction as follows, where C_i^{-j} is the corresponding greater-context when l_{ij} is removed:

$$\begin{gathered}\textit{Semi-Contrapositive Inference: LOO} \\ \exists\, l_{ij} \mid \max_{j} ||\mathcal{F}(C_i) - \mathcal{F}(C_i^{-j})||^2\end{gathered} \tag{5}$$

Both direct and contrapositive inference schemes are relaxing the local prediction requirement of having a large training corpus with local annotation into having training corpus with global annotation, which is easier to obtain, thus, they could be beneficial for new or narrowly focused NLP applications for which localized training data is not widely available. However, it must be noted that the localized prediction inference is only applicable to certain classification problems, where the prediction task can be performed in multiple levels of granularity and the local prediction has a corresponding global prediction task at a greater-context.

3 Transfer Function Model

Since our approach is based on the difference between original and generated greater-contexts, extensive or random modification of the original text might result in meaningless differences, so the key requirements of an ideal transfer function for our approach are:

- **Req. 1:** preserve most of the original content and only make minimal changes.
- **Req. 2:** these minimal changes negate the prediction for the resulting greater-context (makes it belong to the other class).

Any reasonable transfer function that satisfies these requirements may make a suitable model to serve as the core of our proposed approach. As an implementational choice, we adapt the model proposed by Hu et al. (2017) and augmented it with an extra regularization (i.e., conciseness loss, given in Equation 8) to explicitly control for Req. 1.

Our transfer function model incorporates an autoencoder and a discriminator for the global prediction. The autoencoder is initially trained to learn a latent space (z) from the input greater-context instances (c) by minimizing the reconstruction loss, given in Equation (6), where θ_E and θ_G are the encoder and decoder parameters, respectively.

$$\mathcal{L}_{rec}(\theta_E, \theta_G) = \mathbb{E}_{q_E(z|c)}[\log p(c|z, y)] \tag{6}$$

Since the latent space z is learning independent of the class labels ($q_E(z|c)$), it could be used to generate instances of both classes. Therefore, the decoder is trained to learn to generate a greater-context $\hat{c}$, given a desired class label ($p(\hat{c}|z, y)$). Now, if we ask the encoder to generate a rewriting alternative for an input greater-context with a flipped class label ($\neg y = 1 - y$), then the generated output would likely be a modified version of the input with some small changes that make it belong to the opposite class (Req. 2).

However, in order to learn class transference, the decoder also requires a signal to determine how likely the generated output belongs to the desired class. To provide that signal, we incorporate a class discriminator in our transfer function model, which is optimized as given in Equation 7, where θ_D is the discriminator parameter. For simplicity, we only train the discriminator on the input examples (c) and use that to predict the label for generated examples ($\hat{c}$).

$$\mathcal{L}_{\text{disc}}(\theta_D) = \mathbb{E}_C[\log q_D(y|c)] \tag{7}$$

By optimizing for the reconstruction loss, the autoencoder will learn to generate an output similar to the input, which, in part, may satisfy Req. 1. We still want to directly enforce the class transference using minimal changes, which intuitively means the generated transferred version of the input greater-context that belongs to some $Class\ B$ should be as similar as possible to the input greater-context that originally belongs to some $Class\ A$.

Therefore, we also optimize for a *conciseness loss* ($\mathcal{L}_{con}$) as given in Equation 8, where c is the original input greater-context and $\hat{c}$ is the transferred version of it to the other class. It must be noted that, sometimes, the original input and the transferred output might not have the same dimension. For these cases, we slice the bigger vector as the dimension of the smaller vector.

$$\mathcal{L}_{con}(\theta_E, \theta_G) = \mathbb{E}_{q_E(z|c)}[\log p(\hat{c}|z, \neg y)] \quad (8)$$

Equation (9) shows the transfer function's final objective function, where λ_D and λ_C are the balancing parameter.

$$\min_{\theta_G} \mathcal{L}_{gen} = \mathcal{L}_{rec} + \lambda_D \mathcal{L}_{disc} + \lambda_C \mathcal{L}_{con} \quad (9)$$

For some initial training runs, we train the discriminator and autoencoder independently (λ_D = 0); the discriminator will learn to distinguish between classes, and the autoencoder will learn a latent space that could be used to regenerate instances of both classes. At this step, we do not enforce the minimal changes as well (λ_C = 0). After that, for a few more epochs, we jointly train the autoencoder and discriminator ($\lambda_D \neq 0$) to learn the class transference (replace y with $\neg y = 1 - y$) while enforcing the minimal required changes ($\lambda_C \neq 0$). For more details about the hyperparameters and the training process please see Appendix A.

It must be noted that to avoid the fading gradient phenomenon when training on discrete space, the model feeds with Gumbel-Softmax (Jang et al., 2017) representation over the words to provide a continuous approximation and a stronger signal for optimizing the decoder.

4 Experiments

This work argues that inference by contraposition allows for the identification of semantically relevant local features and thus a more accurate local prediction inference when compared to methods based on direct inference. In order to evaluate this hypothesis, we setup an experiment for local prediction inference at word-level from a large training corpus of naturally-occurring and manually-labeled sentences in the sentiment analysis domain (Section 4.1).

We also argue that the contrapositive inference scheme can be more robustly trained on user-generated and weakly-labeled data compared to the alternative direct inference methods. In order to validate this aspect of the contrapositive local inference, we conducted a few experiments for local prediction inference from augmented training corpora of the global prediction in the sentiment analysis, pleonasm detection, and specificity detection problem domains (Section 4.2). In these set of experiments, we compare two implementation of each inference schemes (a simpler and a more complex models), as follows:

Direct:ATN. For the global prediction, we use a BiLSTM classifier and wrap the weights at each time-step into an attention weight using a multiplicative attention layer (Luong et al., 2015). The local feature with the highest attention weight would receive the same class prediction as the global class label.

Direct:RTNL. We use the model proposed by Lei et al. (2016) as the state-of-the-art implementation for the direct inference of the local prediction as the rationales for the global prediction (Equation 3), tuned for each experimental settings (see Appendix A). Their model consist a RCNN generator and encoder to produce rationales and predictions, and uses REINFORCE for generation learning (Sutton et al., 2000).

Contrapositive:LOO. This baseline serves as a light version of the contrapositive inference of the local prediction through examining the global prediction for different perturbation of a greater-context by leaving one of its local features out. The local feature that causes the highest deviation in global prediction would infer the original global class label (Equation 5). Similar to the *Direct:ATN* baseline, we use a BiLSTM classifier for the global prediction task.

Contrapositive:TF. This model serves as an implementation of our proposed contrapositive inference scheme (Equation 4). We use our transfer function model (Section 3) for contrapositive local prediction inference through rewriting and class transference of the greater-context.

4.1 Local Prediction Inference from Naturally-Occurring Labeled Data

In this experiment, we aim to evaluate the different inference schemes on a inference task from sentence (as the global scope) into word(s) (as the smaller local scope), when a large training corpus of global prediction is available. As a result, we opt to run our experiment on **sentiment analysis**, as problem domain with readily available large training corpora of labeled sentences, such as Yelp Polarity Dataset (YPD) (Zhang et al., 2015). In this domain, the global sentiment prediction task is to *predict whether a sentence has a "positive" or "negative" polarity?* Thus, we evaluate the different inference schemes on the local prediction task of inferring *which word(s) of the sentence expresses a strong sentiment?*

To train our transfer function, we use the 105K naturally-occurring polarity-labeled sentences (*positive* and *negative*) of the YPD that contain 30 or fewer words (refer to as YPD-Train). As a benchmark dataset to evaluate the local inference performance of the models, we manually annotated some held-out sentences of YPD with word-level sentiment labels. We followed the same annotation scheme proposed by (Socher et al., 2013) and created a unigram pool from all the words of the test set; then randomly picked a word and annotated it as *very positive*, *positive*, *neutral*, *negative*, or *very negative*. Finally, filtering for sentences that contains a *very positive* or *very negative* words, we collected a set of 860 held-out sentences with word-level labels as our in-house benchmark dataset (refer to as YPD-Test).

Table 1 compares the precision of different inference schemes on predicting the word(s) that express the same sentiment as the sentence in which it is used, based on the gold annotation in the benchmark dataset (YPD-Test). As shown, the contrapositive inference methods, including LOO, are more effective at inferring the sentiment class label of the word(s) than the direct inference alternatives, and our proposed method (*Contrapositive:TF*) substantially outperformed the best direct inference method (*Direct:RTNL*) by more than 14%. This huge performance difference suggests that considering and controlling the contrapositive contribution of local features to the corresponding global prediction can improve the identification of the most semantically contributing local features and thus infer a more accurate local prediction.

Approach	Local Inference Precision
Direct:ATN	58.4%
Direct:RTNL	63.3%
Contrapositive:LOO	66.2%
Contrapositive:TF	**77.5%**

Table 1: Local sentiment prediction precision of different inference schemes, trained on naturally-occurring labeled examples

4.2 Local Prediction Inference from Weakly-Labeled User-Generated Data

The availability of training corpora is always a bottleneck for the training of complex supervised computational models. Our proposed inference scheme already relaxes the requirement of local prediction task from having labeled examples with annotation at smaller scope to having training examples with global annotation, which is usually easier to obtain. Nevertheless, in many resource-constrained NLP problem domains, even training corpora for global prediction are not available. Therefore, in this experiment we investigate whether data augmentation can reconcile the training requirements of different inference schemes by studying two following resource-constrained problem domains:

Pleonasm Detection. This task at local level aims to *find redundant words that are not contributing to the overall meaning of a sentence* (Quinn, 1993; Lehmann, 2005). For example, the word "free" may deem semantically redundant at the presence of the word "gift" in the sentence: "I received a free gift." Like sentiment analysis, this domain has a clear corresponding global classification task: *whether the sentence is "concise" or "verbose"?* However, this domain has a much more limited set of existing resources: NUCLE covers grammatical redundancy (Dahlmeier et al., 2013), and the Semantic Pleonasm Corpus (SPC), a small corpus with pleonasm annotation at word-level (Kashefi et al., 2018), which we use as the benchmark dataset in this experiment.

Specificity Detection. This task at local level aims to *pinpoint the phrases that uniquely relate the sentence to a particular subject* (Li and Nenkova, 2015; Lugini and Litman, 2018). For example, the phrase "bus accident" in the sentence "10 people killed in bus accident in Pakistan", makes it more specific than the phrase "road accident" in the sentence "10 people killed in road accident in Pakistan." It also has a corresponding global prediction task: *whether the sentence conveying a "specific" or "general" piece of information?* However, the existing resources are limited to just a few small corpora (less than 1K sentences) with sentence-level specificity annotation (Louis and

Task	Source	Heuristic Strategy	Example
Sentiment Analysis	Yelp	• ***Positive*** *Label:* four star and above	→ super *generous* portion
		• *Augmented* ***Negative*** *Heuristic:* substitute a positive word (look up in a polarity vocabulary) with an *antonym*	→ super meager portion
		• ***Negative*** *Label:* two star and below	→ it was not a *good* experience
		• *Augmented* ***Positive*** *Heuristic:* substitute a negative word (look up in a polarity vocabulary) with an *antonym*	→ it was not a bad experience
Pleonasm Detection	Yelp	• ***Concise*** *Label:* Yelp tips (short sentences)	→ *delicious* bread
		• *Augmented* ***Verbose*** *Heuristic:* add a *synonym* next to an adjective	→ *delicious +tasty* bread
		• *Augmented* ***Near-Miss Concise*** *Heuristic:* insert a *non-synonym* word next to an adjective, base on language model prediction	→ *delicious +redolent* bread
Specificity Detection	News Headline	• ***Specific*** *Label:* contains more than 2 named entities	→ *Rouhani* wants nuclear deal
		• *Augmented* ***General*** *Heuristic:* substitute a noun with an *hyponym*	→ President wants nuclear deal
		• ***General*** *Label:* contains no named entities	→ 10 killed in *camp*
		• *Augmented* ***Specific*** *Heuristic:* substitute a noun with an *hypernym*	→ 10 killed in death camp

Table 2: Augmentation heuristic strategies for different problem domains

Nenkova, 2012; Louis et al., 2013; Tan and Lee, 2014), and the Interpretable Semantic Textual Similarity (iSTS) dataset, which comprises phrase-level specificity annotation (Agirre et al., 2016), which we use as the benchmark in our experiment.

In addition, we also make a low-resource case for the **sentiment analysis** domain in order to compare the prediction result of our approach with previous experiment, where a sizable training corpus was available for global prediction.

4.2.1 Data Augmentation

In order to augment a sizable dataset to train the inference models, we start by identifying an existing real-world data source that possesses some characteristics that allow us to (weakly) label some of its examples with at least one of the desired classes. For example, the Yelp dataset[2] has a data category called "tips." Since "tips" are very short sentences, they are likely to be *concise*; or Yelp reviews with 4 star and above are likely to carry a *positive* sentiment. Next, we apply some domain-inspired non-label-preserving augmentation heuristics to the instances of one class (e.g., "positive") to generate instances of the opposite class (e.g., "negative"). The details of the heuristics and data augmentation for these domains are discussed elsewhere (Kashefi and Hwa, 2020).

Table 2 summarizes the heuristic strategy we used to augment a training dataset for each problem domain. When evaluating the models for sentiment analysis and pleonasm detection tasks, we use the YPD-Test and SPC, respectively, which are both collected from Yelp. For the specificity detection task, we use the news headline part of the iSTS as the benchmark to evaluate the model. Thus, for data augmentation we use "all the news[3]" corpus, as a data source in a related domain. The resulting augmented training corpora for sentiment analysis, pleonasm detection, and specificity detection tasks are containing 105k (the same size as YPD-Train in previous experiment), 160K and 110K user-generated examples, respectively.

4.2.2 Results

Table 3 compares the local prediction inference precision of the different inference schemes across different problem domains, when trained on noisy user-generated corpora, based on the gold annotation in the benchmark datasets.

We can observer that the contrapositive local inference schemes, including *Contrapositive:LOO*, outperform the direct inference schemes in all problem domains and the local prediction precision of our proposed approach (*Contrapositive:TF*) is significantly higher than all other inference alternatives. As expected, the local prediction inference precision of *Contrapositive:TF* was slightly lower than its corresponding performance in the previous experiment, where models were trained on

[2] www.yelp.com/dataset/challenge

[3] www.kaggle.com/snapcrack/all-the-news

Inference Approach	**Local Inference Precision**		
	Sentiment Analysis	Pleonasm Detection	Specificity Detection
Direct:ATN	54.3%	24.3%	20.5%
Direct:RTNL	55.6%	30.4%	28.1%
Contrapositive:LOO	59.8%	41.1%	45.5%
Contrapositive:TF	**74.2%**	**70.3%**	**69.5%**

Table 3: Localized prediction precision of the approaches on different problem domains, trained on weakly-labeled augmented datasets

manually-labeled real examples (74.2% compared to 77.5%). However, the performance gap between our proposed model and other models are getting larger when trained on noisy data (e.g., 14% difference with *Direct:RTNL* on real data compared to 19% difference with noisy data).

These results may suggest that the contrapositive inference is more noise-tolerant than the alternative methods. In addition, the impact of data augmentation on the performance of our approach is not significant, so it could be beneficial for providing training requirements of our approach and making it applicable to many low-resource NLP problems.

5 Discussion

Results from our studies suggest that enforcing the contraposition constraint between the local prediction and its corresponding global prediction can reveal their semantic relationship more robustly, and thus, lead to a a more accurate local class prediction inference, compared to direct inference alternatives.

In the direct inference schemes, there are many cases that semantically irrelevant words, such as "punctuation marks", are the heavy contributors to the global prediction. These observations and the performance of the *Direct:Atn* baseline confirm the prior work's suggestions that attention weight may not correlate with the semantics of the prediction task. In addition, the *Direct:Rationale* baseline operates through finding a smaller local text span (rationale) that can replace the greater-context in the global prediction. Intuitively, rationales with shorter text span (e.g., words or phrases) are likely to provide a noisy and unstable solution for the global problem, as we observed in our experiments.

The *Contrapositive:LOO* baseline adopts a simple approach to apply a relaxed version of the contraposition constraint, by removing the local rationale from the global problem's context. As observed, a global context, after removing a small text span from it, would still be large and semantically expressive enough for a reliable global prediction.

Since our proposed *Contrapositive:TF* approach operates by replacing a local rationale of a global context with another (adversarial) rationale, the span size of the greater-context remains relatively intact, which makes the prediction inference more robust. The transfer function also operates by disentangling the semantics from the surface representation and applying local contrapositive semantic perturbation to the global context, which makes it more robust to noise and irrelevant local features. For example, a punctuation mark might influence a class prediction for a sentence, however, replacing it with another punctuation mark will not make the sentence belong to another class, so it is not a *semantically* contributing word.

Furthermore, evaluation results suggest that the performance impact of training a transfer function on weakly-labeled user-generated data is not significant, so data augmentation may facilitate the training of the transfer function and make our approach applicable to a variety of resource-constrained NLP problem domains. However, in our experiments, we observed that the quality of augmented examples is a key-factor for training an appropriate model. We found that both global and local classification performance declined significantly when we trained models on the corpora that are augmented with simple heuristics, which generate obvious examples for different classes.

This should be noted that our proposed contrapositive approach is only applicable to prediction tasks that can be performed in multiple levels of granularity, where the local prediction has a corresponding global prediction task. In addition, applying it to more general cases, such as "multi-class" classification, might pose some challenges.

6 Conclusion

This paper presents a local contrapositive inference scheme that is only informed by corresponding global class predictions at a greater contexts. The inference scheme is implemented as a transfer function that learns to transform a context from one class to another. Training such a transfer function, requires a large corpus of labeled sentences, which may not be available for many low-resource problem domains. Our work demonstrates the robustness of the proposed approach when coupled with appropriate data augmentation methods and its applicability as a solution to resource-constrained local prediction tasks.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments. This material is based upon work supported by the National Science Foundation under Grant Number 1735752.

References

Eneko Agirre, Aitor Gonzalez-Agirre, Inigo Lopez-Gazpio, Montse Maritxalar, German Rigau, and Larraitz Uria. 2016. SemEval-2016 Task 2: Interpretable Semantic Textual Similarity. In *SemEval*.

Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. 2018. Deriving Machine Attention from Human Rationales. In *EMNLP*.

Samuel R Bowman, Vilnis Luke, Vinyals Oriol, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating Sentences from a Continuous Space. In *CoNLL*.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English. In *Workshop on Innovative Use of NLP for Building Educational Applications*.

Mengnan Du, Ninghao Liu, Fan Yang, and Xia Hu. 2019. Learning Credible Deep Neural Networks with Rationale Regularization. In *ICDM*.

Upol Ehsan, Brent Harrison, Larry Chan, and Mark O Riedl. 2018. Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations. In *AIES*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward Controlled Generation of Text. In *ICML*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not Explanation. In *EMNLP*.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical Reparameterization With Gumbel-Softmax. In *ICLR*.

Omid Kashefi and Rebecca Hwa. 2020. Quantifying the Evaluation of Heuristic Methods for Textual Data Augmentation. In *EMNLP Workshop on Noisy User-generated Text (W-NUT)*.

Omid Kashefi, Andrew T Lucas, and Rebecca Hwa. 2018. Semantic Pleonasm Detection. In *NAACL*.

Christian Lehmann. 2005. Pleonasm and hypercharacterisation. In *Yearbook of Morphology*, pages 119–154. Springer, Dordrecht.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing Neural Predictions. In *EMNLP*.

Junyi Jessy Li and Ani Nenkova. 2015. Fast and Accurate Prediction of Sentence Specificity. In *AAAI*.

Annie Louis, Stefanie Dipper, Heike Zinsmeister, and Bonnie Webber. 2013. A corpus of science journalism for analyzing writing quality. *Dialogue and Discourse*, 4(2):87–117.

Annie Louis and Ani Nenkova. 2012. A Corpus of General and Specific Sentences from News. In *LREC*.

Luca Lugini and Diane Litman. 2018. Predicting Specificity in Classroom Discussion. In *Workshop on Innovative Use of NLP for Building Educational Applications*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *EMNLP*.

Iain J Marshall, Joë Kuiper, and Byron C Wallace. 2016. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association*, 23(1):193–201.

Pramod K. Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the model understand the question? In *ACL*.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style Transfer Through Back-Translation. In *ACL*.

Arthur. Quinn. 1993. *Figures of speech : 60 ways to turn a phrase*. Psychology Press.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *KDD*.

Tianxiao Shen, Tao Lei, Regina Barzilay, Tommi Jaakkola, and Mit Csail. 2017. Style Transfer from Non-Parallel Text by Cross-Alignment. In *NIPS*.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *EMNLP*.

Julia Strout, Ye Zhang, and Raymond J. Mooney. 2019. Do Human Rationales Improve Machine Explanations? In *BlackboxNLP Workshop at ACL*.

Richard S Sutton, David Mcallester, Satinder Singh, and Yishay Mansour. 2000. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *NIPS*.

Chenhao Tan and Lillian Lee. 2014. A Corpus of Sentence-level Revisions in Academic Writing: A Step towards Understanding Statement Strength in Communication. In *ACL*.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P. Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised Text Style Transfer using Language Models as Discriminators. In *NIPS*.

Omar F Zaidan, Jason Eisner, and Christine D Piatko. 2007. Using "Annotator Rationales" to Improve Machine Learning for Text Categorization. In *NAACL*.

Xiang Zhang, Junbo Zhao, and Yann Lecun. 2015. Character-level Convolutional Networks for Text Classification. In *NIPS*.

Ye Zhang, Iain Marshall, and Byron C Wallace. 2016. Rationale-Augmented Convolutional Neural Networks for Text Classification. In *EMNLP*.

Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen. 2019. Style Transfer as Unsupervised Machine Translation. In *AAAI*.

A Hyperparameters

Direct:Atn: attention weight over single words.

Contrapositive:LOO: single words as local text span.

Contrapositive:TF:

- Pre-training
  - #epochs = 10
  - $\lambda_D = .0$
  - $\lambda_C = .0$
- Joint-training
  - #epochs = 2
  - $\lambda_D = 1e - 1$
  - $\lambda_C = 1e - 2$

Direct:Rationale:

- Experiment 1
  - Embedding: GloVe + Yelp
  - $\lambda_1 = 1e - 2$
  - $\lambda_2 = 2\lambda_1$
- Experiment 2
  - Sentiment Analysis Task
    - Embedding: GloVe + Yelp
    - $\lambda_1 = 3e - 2$
    - $\lambda_2 = 2\lambda_1$
  - Pleonasm Detection Task
    - Embedding: GloVe + Yelp
    - $\lambda_1 = 5e - 2$
    - $\lambda_2 = 2\lambda_1$
  - Specificity Detection Task
    - Embedding: GloVe + “all the news”
    - $\lambda_1 = 1e - 2$
    - $\lambda_2 = 2\lambda_1$