# Causal Direction of Data Collection Matters: Implications of Causal and Anticausal Learning for NLP

**Zhijing Jin[1,2*], Julius von Kügelgen[1,3*], Jingwei Ni[4], Tejas Vaidhya[5], Ayush Kaushal[5], Mrinmaya Sachan[2] and Bernhard Schölkopf[1,2]**
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany,
[2]ETH Zürich, [3]University of Cambridge, [4]University College London, [5]IIT Kharagpur
{zjin,jvk,bs}@tue.mpg.de, ucabjni@ucl.ac.uk,
{ayushkaushal, iamtejasvaidhya}@iitkgp.ac.in, msachan@ethz.ch

## Abstract

The principle of independent causal mechanisms (ICM) states that generative processes of real world data consist of independent modules which do not influence or inform each other. While this idea has led to fruitful developments in the field of causal inference, it is not widely-known in the NLP community. In this work, we argue that the causal direction of the data collection process bears nontrivial implications that can explain a number of published NLP findings, such as differences in semi-supervised learning (SSL) and domain adaptation (DA) performance across different settings. We categorize common NLP tasks according to their causal direction and empirically assay the validity of the ICM principle for text data using minimum description length. We conduct an extensive meta-analysis of over 100 published SSL and 30 DA studies, and find that the results are consistent with our expectations based on causal insights. This work presents the first attempt to analyze the ICM principle in NLP, and provides constructive suggestions for future modeling choices.[1]

## 1 Introduction

NLP practitioners typically do not pay great attention to the causal direction of the data collection process. As a motivating example, consider the case of collecting a dataset to train a machine translation (MT) model to translate from English (En) to Spanish (Es): it is common practice to mix all available En-Es sentence pairs together and train the model on the entire pooled data set (Bahdanau et al., 2015; Cho et al., 2014). However, such mixed corpora actually consist of two distinct types of data: (i) sentences that originated in English and have been translated (by human translators) into Spanish (En→Es); and (ii) sentences that originated in Spanish and have subsequently been translated into English (Es→En).[2]

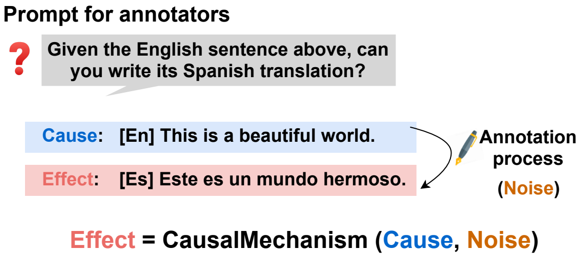
Prompt for annotators
Given the English sentence above, can you write its Spanish translation?
Cause: [En] This is a beautiful world.
Effect: [Es] Este es un mundo hermoso.
Annotation process
(Noise)
Effect = CausalMechanism (Cause, Noise)

Figure 1: Annotation process for NLP data: the random variable that exists first is typically the cause (e.g., a given prompt), and the one generated afterwards is typically the effect (e.g., the annotated answer).

Intuitively, these two subsets are qualitatively different, and an increasing number of observations by the NLP community indeed suggests that they exhibit different properties (Freitag et al., 2019; Edunov et al., 2020; Riley et al., 2020; Shen et al., 2021). In the case of MT, for example, researchers find that training models on each of these two types of data separately leads to different test performance, as well as different performance improvement by semi-supervised learning (SSL) (Bogoychev and Sennrich, 2019; Graham et al., 2020; Edunov et al., 2020). Motivated by this observation that the data collection process seems to matter for model performance, in this work, we provide an explanation of this phenomenon from the perspective of causality (Pearl, 2009; Peters et al., 2017).

First, we introduce the notion of the *causal direction* for a given NLP task, see Fig. 1 for an example. Throughout, we denote the input of a learning task by $X$ and the output which is to be predicted by $Y$. If, during the data collection process, $X$ is generated first, and then $Y$ is collected based on $X$ (e.g., through annotation), we say that $X$ causes $Y$, and denote this by $X \rightarrow Y$. If, on the other hand, $Y$ is

[*] Equal contribution.

[1] The codes are at https://github.com/zhijing-jin/icm4nlp.

[2] There is, in principle, a third option: both could be translations from a third language, but this occurs less frequently.

*Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9499–9513
November 7–11, 2021. ©2021 Association for Computational Linguistics

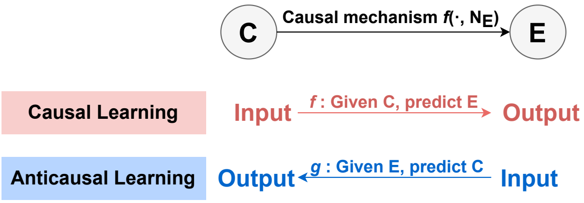

Figure 2: *(Top)* A causal graph $C \to E$, where $C$ is the cause and $E$ is the effect. The function $f(\cdot, N_E)$ denotes the causal process, or mechanism, $P_{E|C}$ by which the effect $E$ is generated from $C$ and unobserved noise $N_E$. *(Bottom)* Based on whether the direction of prediction aligns with the direction of causation or not, we distinguish two types of tasks: (i) causal learning, i.e., predicting the effect from the cause; and (ii) anticausal learning, i.e., predicting the cause from the effect.

| Category | Example NLP Tasks |
|---|---|
| **Causal learning** | Summarization, parsing, tagging, data-to-text generation, information extraction |
| **Anticausal learning** | Author attribute classification, review sentiment classification |
| **Other/mixed (depending on data collection)** | Machine translation, question answering, question generation, intent classification |

Table 1: Classification of typical NLP tasks into causal (where the model takes the cause as input and predicts the effect), and anticausal (where the model takes the effect as input and predicts the cause) learning problems, as well as other tasks which do not have a clear causal interpretation of the data collection process, or where a mixture of both types of data is typically used.

generated first, and then $X$ is collected based on $Y$, we say that $Y$ causes $X$ ($Y \to X$).[3]

Based on whether the direction of prediction aligns with the causal direction of the data collection process or not, Schölkopf et al. (2012) categorize these types of tasks as *causal learning* ($X \to Y$), or *anticausal learning* ($Y \to X$), respectively; see Fig. 2 for an illustration. In the context of our motivating MT example this means that, if the goal is to translate from English ($X =$ En) into Spanish ($Y =$ Es), training *only* on subset (i) of the data consisting of En→Es pairs corresponds to *causal learning* ($X \to Y$), whereas training *only* on subset (ii) consisting of Es→En pairs is categorised as *anticausal learning* ($Y \to X$).

Based on the principle of independent causal mechanisms (ICM) (Janzing and Schölkopf, 2010; Peters et al., 2017), it has been hypothesized that the causal direction of data collection (i.e., whether a given NLP learning task can be classified as causal or anticausal) has implications for the effectiveness of commonly used techniques such as SSL and domain adaptation (DA) (Schölkopf et al., 2012). We will argue that this can explain performance differences reported by the NLP community across different data collection processes and tasks. In particular, we make the following contributions:

1. We categorize a number of common NLP tasks according to the causal direction of the underlying data collection process (§ 2).
2. We review the ICM principle and its implications for common techniques of using unlabelled data such as SSL and DA in the context of causal and anticausal NLP tasks (§ 3).
3. We empirically assay the validity of ICM for NLP data using minimum description length in a machine translation setting (§ 4).
4. We verify experimentally and through a meta-study of over respectively 100 (SSL) and 30 (DA) published findings that the difference in SSL (§ 5) and domain adaptation (DA) (§ 6) performance on causal vs anticausal datasets reported in the literature is consistent with what is predicted by the ICM principle.
5. We make suggestions on how to use findings in this paper for future work in NLP (§ 7).

## 2 Categorization of Common NLP Tasks into Causal and Anticausal Learning

We start by categorizing common NLP tasks which use an input variable $X$ to predict a target or output variable $Y$ into causal learning ($X \to Y$), anticausal learning ($Y \to X$), and other tasks that do not have a clear underlying causal direction, or which typically rely on mixed (causal and anticausal) types of data, as summarised in Tab. 1.

Key to this categorization is determining whether the input $X$ corresponds to the cause or the effect in the data collection process. As illustrated in Fig. 1, if the input $X$ and output $Y$ are generated at two different time steps, then the variable that is generated first is typically the cause, and the other that is subsequently generated is typically the effect, provided it is generated based on the previous one (rather than, say, on a common confounder that causes both variables). If $X$ and $Y$ are generated jointly, then we need to distinguish based on the underlying generative process whether one of the two variables is causing the other variable.

[3] This corresponds to an *interventional* notion of causation: if one were to manipulate the cause, the annotation process would lead to a potentially different effect. A manipulation of the effect, in contrast, would not change the cause.

**Learning Effect from Cause (Causal Learning)** Causal ($X \to Y$) NLP tasks typically aim to predict a post-hoc generated human annotation (i.e., the target $Y$ is the effect) from a given input $X$ (the cause). Examples include: summarization (*article→summary*) where the goal is to produce a summary $Y$ of a given input text $X$; parsing and tagging (*text→linguists' annotated structure*) where the goal is to predict an annotated syntactic structure $Y$ of a given input sentence $X$; data-to-text generation (*data→description*) where the goal is to produce a textual description $Y$ of a set of structured input data $X$; and information extraction (*text→entities/relations/etc*) where the goal is to extract structured information from a given text.

**Learning Cause from Effect (Anticausal Learning)** Anticausal ($Y \to X$) NLP tasks typically aim to predict or infer some latent target property $Y$ such as an unobserved prompt from an observed input $X$ which takes the form of one of its effects. Typical anticausal NLP learning problems include, for example, author attribute identification (*author attribute→text*) where the goal is to predict some unobserved attribute $Y$ of the writer of a given text snippet $X$; and review sentiment classification (*sentiment→review text*) where the goal is to predict the latent sentiment $Y$ that caused an author to write a particular review $X$.

**Other/Mixed** Some tasks can be categorized as either causal or anticausal, depending on how exactly the data is collected. In § 1, we discussed the example of MT where different types of (causal and anticausal) data are typically mixed. Another example is the task of intent classification: if the *same* author reveals their intent before the writing (i.e., *intent→text*), it can be viewed as an anticausal learning task; if, on the other hand, the data is annotated by *other* people who are not the original author (i.e., *text→annotated intent*), it can be viewed as a causal learning task. A similar reasoning applies to question answering and generation tasks which respectively aim to provide an answer to a given question, or vice versa: if first a piece of informative text is selected and annotators are then asked to come up with a corresponding question (*answer→question*) as, e.g., in the SQuAD dataset (Rajpurkar et al., 2016), then question answering is an anticausal and question generation a causal learning task; if, on the other hand, a question such as a search query is selected first and subsequently an answer is provided (*question→answer*) as, e.g., in the Natural Questions dataset (Kwiatkowski et al., 2019), then question answering is a causal and question generation an anticausal learning task. Often, multiple such datasets are combined without regard for their causal direction.

## 3 Implications of ICM for Causal and Anticausal Learning Problems

Whether we are in a causal or anticausal learning scenario has important implications for semi-supervised learning (SSL) and domain adaptation (DA) (Schölkopf et al., 2012; Sgouritsa et al., 2015; Zhang et al., 2013, 2015; Gong et al., 2016; von Kügelgen et al., 2019, 2020), which are techniques also commonly used in NLP. These implications are derived from the principle of independent causal mechanisms (ICM) (Schölkopf et al., 2012; Lemeire and Dirkx, 2006) which states that "*the causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other*" (Peters et al., 2017).

In the bivariate case, this amount to a type of independence assumption between the distribution $P_C$ of the cause $C$, and the causal process, or mechanism, $P_{E|C}$ that generates the effect from the cause. For example, for a question answering task, the generative process $P_C$ by which one person comes up with a question $C$ is "independent" of the process $P_{E|C}$ by which another person produces an answer $E$ for question $C$.[4]

Here, "independent" is not meant in the sense of *statistical* independence of random variables, but rather as *independence at the level of generative processes or distributions* in the sense that $P_C$ and $P_{E|C}$ *do not share information* (the person asking the question and the one answering may not know each other) and *can be manipulated independently of each other* (we can swap either of the two for another participant without the other one being influenced by this). Crucially, this type of independence is generally violated in the opposite, i.e., *anticausal*, direction: $P_E$ and $P_{C|E}$ may share information and change dependently (Daniušis et al., 2010; Janzing et al., 2012). This has two important implications for common learning tasks (Schölkopf et al., 2012) which are illustrated in Fig. 3.

[4] The validity of this is meant in an approximate sense, and one can imagine settings where it is questionable. E.g., if the person asking the question has prior knowledge of the respondent (e.g., in a classroom setting), then she might adjust the question accordingly which would violate the assumption.

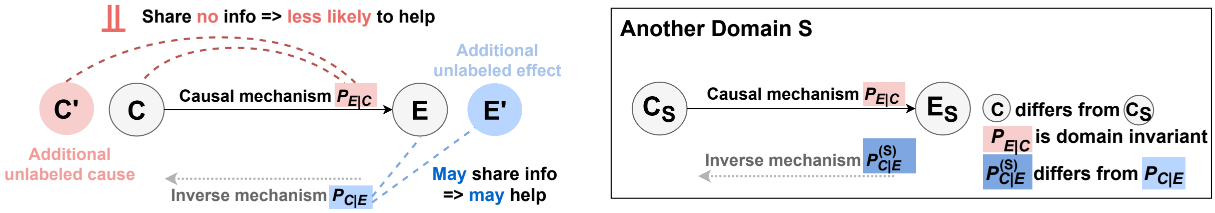
Share no info => less likely to help
Additional unlabeled effect
C'
C
Causal mechanism $P_{E|C}$
E
E'
Additional unlabeled cause
Inverse mechanism $P_{C|E}$
May share info => may help
Another Domain S
$C_S$
Causal mechanism $P_{E|C}$
$E_S$
C differs from $C_S$
$P_{E|C}$ is domain invariant
Inverse mechanism $P^{(S)}_{C|E}$
$P^{(S)}_{C|E}$ differs from $P_{C|E}$

Figure 3: The ICM principle assumes that *the generative process $P_C$ of the cause $C$ is independent of the causal mechanism $P_{E|C}$*: the two distributions share no information and each may be changed or manipulated without affecting the other. In the anticausal direction, on the other hand, the effect distribution $P_E$ *is (in the generic case) not independent of the inverse mechanism $P_{C|E}$*: they may share information and change dependently. *(Left)* SSL, which aims to improve an estimate of the target conditional $P_{Y|X}$ given additional unlabelled input data from $P_X$, should therefore not help for causal learning ($X \to Y$), but may help in the anticausal direction ($Y \to X$). *(Right)* DA, which aims to adapt a model of $P_{Y|X}$ from a source domain to a target domain (e.g., fine-tuning on a smaller dataset), should work better for causal learning settings where a change in $P_C$ is not expected to lead to a change in the mechanism $P_{E|C}$, whereas in the anticausal direction $P_E$ and $P_{C|E}$ may change in a *dependent* manner.

**Implications of ICM for SSL** First, if $P_C$ shares no information with $P_{E|C}$, SSL—where one has additional unlabelled input data from $P_X$ and aims to improve an estimate of the target conditional $P_{Y|X}$—should not work in the causal direction ($X \to Y$), but may work in the anticausal direction ($Y \to X$), as $P_E$ and $P_{C|E}$ may share information. Causal NLP tasks should thus be less likely to show improvements over a supervised baseline when using SSL than anticausal tasks.

**Implications of ICM for DA** Second, according to the ICM principle, the causal mechanism $P_{E|C}$ should be invariant to changes in the cause distribution $P_C$, so domain—specifically, covariate shift (Shimodaira, 2000; Sugiyama and Kawanabe, 2012)—adaptation, where $P_X$ changes but $P_{Y|X}$ is assumed to stay invariant, should work in the causal direction, but not necessarily in the anticausal direction. Hence, DA should be easier for causal NLP tasks than for anticausal NLP tasks.

## 4 Investigating the Validity of ICM for NLP Data Using MDL

Traditionally, the ICM principle is thought of in the context of *physical* processes or mechanisms, rather than *social* or *linguistic* ones such as language. Since ICM amounts to an independence assumption that—while well motivated in principle—may not always hold in practice,[5] we now assay its validity on NLP data.

Recall, that ICM postulates a type of independence between $P_C$ and $P_{E|C}$. One way to formalize this uses Kolmogorov complexity $K(\cdot)$ as a measure of algorithmic information, which can be understood as the length of the shortest program that computes a particular algorithmic object such as a distribution or a function (Solomonoff, 1964; Kolmogorov, 1965). ICM then reads (Janzing and Schölkopf, 2010):[6]

$$\begin{aligned} K(P_{C,E}) &\overset{+}{=} K(P_C) + K(P_{E|C}) \\ &\overset{+}{\leq} K(P_E) + K(P_{C|E})\,. \end{aligned} \tag{1}$$

In other words, the shortest description of the joint distribution $P_{C,E}$ corresponds to describing $P_C$ and $P_{E|C}$ separately (i.e., they share no information), whereas there may be redundant (shared) information in the non-causal direction such that a separate description of $P_E$ and $P_{C|E}$ will generally be longer than that of the joint distribution $P_{C,E}$.

### 4.1 Estimation by MDL

Since Kolmogorov complexity is not computable (Li et al., 2008), we adopt a commonly used proxy, the minimum description length (MDL) (Grünwald, 2007), to test the applicability of ICM for NLP data. Given an input, such as a collection of observations $\{(c_i, e_i)\}_{i=1}^n \sim P_{C,E}$, MDL returns the shortest codelength (in bits) needed to compress the input, as well as the parameters needed to decompress it. We use MDL to approximate (1) as follows:

$$\begin{aligned} \mathrm{MDL}(\mathbf{c}_{1:n}, \mathbf{e}_{1:n}) &= \mathrm{MDL}(\mathbf{c}_{1:n}) + \mathrm{MDL}(\mathbf{e}_{1:n}|\mathbf{c}_{1:n}) \\ &\leq \mathrm{MDL}(\mathbf{e}_{1:n}) + \mathrm{MDL}(\mathbf{c}_{1:n}|\mathbf{e}_{1:n}), \end{aligned} \tag{2}$$

where $\mathrm{MDL}(\cdot|\cdot)$ denotes a conditional compression where the second argument is treated as "free parameters" which do not count towards the compression length of the first argument. Eq. (2) can thus

[5] E.g., due to confounding influences from unobserved variables, or mechanisms which have co-evolved to be dependent

[6] Here, $\overset{+}{=}$ and $\overset{+}{\leq}$ hold up a constant due to the choice of a Turing machine in the definition of algorithmic information.

be interpreted as a comparison between two ways of compressing the same data $(\mathbf{c}_{1:n}, \mathbf{e}_{1:n})$: either we first compress $\mathbf{c}_{1:n}$ and then compress $\mathbf{e}_{1:n}$ conditional on $\mathbf{c}_{1:n}$, or vice versa. According to the ICM principle, the first way should tend to be more "concise" than the second.

### 4.2 Calculating MDL Using Machine Translation as a Case Study

To empirically assess the validity of ICM for NLP data using MDL as a proxy, we turn to MT as a case study. We choose MT because the input and output spaces of MT are relatively symmetric, as opposed to other NLP tasks such as text classification where the input space is sequences, but the output space is a small set of labels.

There are only very few studies which calculate MDL on NLP data, so we extend the method of Voita and Titov (2020) to calculate MDL using online codes (Rissanen, 1984) for deep learning tasks (Blier and Ollivier, 2018). Since the original calculation method for MDL by Voita and Titov (2020) was developed for classification, we extend it to sequence-to-sequence (Seq2Seq) generation. Specifically, given a translation dataset $D = \{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_n, \mathbf{y}_n)\}$ of $n$ pairs of sentences $\mathbf{x}_i$ with translation $\mathbf{y}_i$, denote the size of the vocabulary of the source language by $V_x$, and the size of the vocabulary of the target language by $V_y$. In order to assess whether (2) holds, we need to calculate four different terms: two marginal terms $\mathrm{MDL}(\mathbf{x}_{1:n})$ and $\mathrm{MDL}(\mathbf{y}_{1:n})$, and two conditional terms $\mathrm{MDL}(\mathbf{y}_{1:n}|\mathbf{x}_{1:n})$ and $\mathrm{MDL}(\mathbf{x}_{1:n}|\mathbf{y}_{1:n})$.

**Codelength of the Conditional Terms** To calculate the codelength of the two conditional terms, we extend the method of Voita and Titov (2020) from classification to Seq2Seq generation. Following the setting of Voita and Titov (2020), we break the dataset $D$ into 10 disjoint subsets with increasing sizes and denote the end index of each subset as $t_i$.[7] We then estimate $\mathrm{MDL}(\mathbf{y}_{1:n}|\mathbf{x}_{1:n})$ as

$$\begin{aligned}\widehat{\mathrm{MDL}}(\mathbf{y}_{1:n}|\mathbf{x}_{1:n}) &= \textstyle\sum_{i=1}^{t_1} \mathrm{length}(\mathbf{y}_i) \cdot \log_2 V_y \\ &- \textstyle\sum_{i=1}^{n-1} \log_2 p_{\theta_i}(\mathbf{y}_{1+t_i:t_{i+1}}|\mathbf{x}_{1+t_i:t_{i+1}}) \;, \end{aligned} \tag{3}$$

where $\mathrm{length}(\mathbf{y}_i)$ refers to the number of tokens in the sequence $\mathbf{y}_i$, $\theta_i$ are the parameters of a translation model $h_i$ trained on the first $t_i$ data points, and $\mathbf{seq}_{\mathrm{idx}_1:\mathrm{idx}_2}$ refers to the set of sequences from

| Dataset | Size | Note |
|---|---|---|
| En→Es | 81K | Original English, Translated Spanish |
| Es→En | 81K | Original Spanish, Translated English |
| En→Fr | 16K | Original English, Translated French |
| Fr→En | 16K | Original French, Translated English |
| Es→Fr | 15K | Original Spanish, Translated French |
| Fr→Es | 15K | Original French, Translated Spanish |

Table 2: Details of the CausalMT corpus.

the $\mathrm{idx}_1$-th to the $\mathrm{idx}_2$-th sample in the dataset $D$, where $\mathbf{seq} \in \{\mathbf{x}, \mathbf{y}\}$ and $\mathrm{idx}_i \in \{1, \dots, n\}$. Similarly, when calculating $\mathrm{MDL}(\mathbf{x}_{1:n}|\mathbf{y}_{1:n})$, we simply swap the roles of $\mathbf{x}$ and $\mathbf{y}$.

**Codelength of the Marginal Terms** When calculating the two marginal terms, $\mathrm{MDL}(\mathbf{x}_{1:n})$ and $\mathrm{MDL}(\mathbf{y}_{1:n})$, we make two changes from the above calculation of conditional terms: first, we replace the *translation models* $h_i$ with *language models*; second, we remove the conditional distribution. That is, we calculate $\mathrm{MDL}(\mathbf{x}_{1:n})$ as

$$\begin{aligned}\widehat{\mathrm{MDL}}(\mathbf{x}_{1:n}) &= \textstyle\sum_{i=1}^{t_1} \mathrm{length}(\mathbf{x}_i) \cdot \log_2 V_x \\ &- \textstyle\sum_{i=1}^{n-1} \log_2 p_{\theta_i}(\mathbf{x}_{1+t_i:t_{i+1}}) \;, \end{aligned} \tag{4}$$

where $\theta_i$ are the parameters of a language model $h_i$ trained on the first $t_i$ data points. We apply the same method to calculate $\mathrm{MDL}(\mathbf{y}_{1:n})$.

For the language model, we use GPT2 (Radford et al., 2019), and for the translation model, we use the Marian neural machine translation model (Junczys-Dowmunt et al., 2018) trained on the OPUS Corpus (Tiedemann and Nygaard, 2004). For fair comparison, all models adopt the transformer architecture (Vaswani et al., 2017), and have roughly the same number of parameters. See Appendix B for more experimental details.

### 4.3 CausalMT Corpus

For our MDL experiment, we need datasets for which the causal direction of data collection is known, i.e., for which we have ground-truth annotation of which text is the original and which is a translation, instead of a mixture of both. Since existing MT corpora do not have this property as discussed in § 1, we curate our own corpus, which we call the CausalMT corpus.

Specifically, we consider the existing MT dataset WMT'19,[8] and identify some subsets that have a clear notion of causality. The subsets we use are the EuroParl (Koehn, 2005) and Global Voices

[7] The sizes of the 10 subsets are 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.25, 12.5, 25, and 50 percent of the dataset size, respectively. E.g., $t_1 = 0.1\%n, t_2 = (0.1\% + 0.2\%)n, \dots$.

[8] Link to WMT'19.

| Data (X→Y) | MDL(X) | MDL(Y) | MDL(Y\|X) | MDL(X\|Y) | MDL(X)+MDL(Y\|X) vs. MDL(Y)+MDL(X\|Y) |
|---|---|---|---|---|---|
| En→Es | 46.54 | 105.99 | 2033.95 | 2320.93 | 2080.49 < 2426.92 |
| Es→En | 113.42 | 55.79 | 3289.99 | 3534.09 | 3403.41 < 3589.88 |
| En→Fr | 20.54 | 53.83 | 503.78 | 535.88 | 524.32 < 589.71 |
| Fr→En | 53.83 | 21.6 | 705.28 | 681.12 | 759.11 > 702.72 |
| Es→Fr | 58.26 | 55.66 | 701.04 | 755.5 | 759.30 < 811.16 |
| Fr→Es | 56.14 | 54.34 | 665.26 | 706.53 | 721.40 < 760.87 |

Table 3: Codelength (in kbits) of $\mathrm{MDL}(X)$, $\mathrm{MDL}(Y)$, $\mathrm{MDL}(Y|X)$, and $\mathrm{MDL}(X|Y)$ on six CausalMT datasets.

translation corpora.[9] For EuroParl, each text has meta information such as the speaker's language; for Global Voices, each text has meta information about whether it is translated or not. We regard text that is in the same language as the speaker's native language in EuroParl (and non-translated text in Global Voices) as the original (i.e., the cause). We then retrieve a corresponding effect by using the cause text to match the parallel pairs in the processed dataset. In this way, we compile six translation datasets with clear causal direction as summarized in Tab. 2. For each dataset, we use 1K samples each as test and validation sets, and use the rest for training.

### 4.4 Results

The results of our MDL experiment on the six CausalMT datasets are summarised in Tab. 3. If ICM holds, we expect the sum of codelengths to be smaller for the causal direction than for the anticausal one, see (2). As can be seen from the last column, this is the case for five out of the six datasets. For example, on one of the largest datasets (En→Es), the MDL difference is 346 kbits.[10]

Comparing the dataset sizes in Tab. 2 and results in Tab. 3, we observe that the absolute MDL values are roughly proportional to dataset size, but other factors such as language and task complexity also play a role. This is inherent to the nature of MDL being the sum of codelengths of the model and of the data given the model. Since we use equally-sized datasets for each language pair in the CausalMT corpus (i.e., in both the $X \to Y$ and $Y \to X$ directions, see Tab. 2), numbers for the same language pair in Tab. 3, including the most important column "MDL(X)+MDL(Y|X) vs. MDL(Y)+MDL(X|Y)", form a valid comparison. That is, En&Es experiments are comparable within themselves, so are the other language pairs.

For some of the smaller differences in the last column in Tab. 3, and, in particular the reversed inequality in row 4, a potential explanation may be the relatively small dataset size, as well as the fact that text data may be confounded (e.g., through shared grammar and semantics).

## 5 SSL for Causal vs. Anticausal Models

In semi-supervised learning (SSL), we are given a typically-small set of $k$ labeled observations $D_L = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \dots, (\boldsymbol{x}_k, \boldsymbol{y}_k)\}$, and a typically-large set of $m$ unlabeled observations of the input $D_U = \{\boldsymbol{x}_1^{(u)}, \dots, \boldsymbol{x}_m^{(u)}\}$. SSL then aims to use the additional information about the input distribution $P_X$ from the unlabeled dataset $D_U$ to improve a model of $P_{Y|X}$ learned on the labeled dataset $D_L$.

As explained in § 3, SSL should only work for anticausal (or confounded) learning tasks, according to the ICM principle. Schölkopf et al. (2012) have observed this trend on a number of classification and regression tasks on small-scale numerical inputs, such as predicting Boston housing prices from quantifiable neighborhood features (causal learning), or breast cancer from lab statistics (anticausal learning). However, there exist no studies investigating the implications of ICM for SSL on NLP data, which is of a more complex nature due to the high dimensionality of the input and output spaces, as well as potentially large confounding. In the following, we use a sequence-to-sequence decipherment experiment (§ 5.1) and a meta-study of existing literature (§ 5.2) to showcase that the same phenomenon also occurs in NLP.

### 5.1 Decipherment Experiment

To have control over causal direction of the data collection process, we use a synthetic decipherment dataset to test the difference in SSL improvement between causal and anticausal learning tasks.

**Dataset** We create a synthetic dataset of encrypted sequences. Specifically, we (i) adopt a monolingual English corpus (for which we use the English corpus of the En→Es in the CausalMT dataset, for

[9] Link to Global Voices.

[10] As far as we know, determining statistical significance in the investigated setting remains an open problem. While, in theory, one may use information entropy to estimate it, in practice, this may be inaccurate since (i) MDL is only a proxy for algorithmic information; and (ii) ICM may not hold exactly, but only approximately. We evaluate on six different datasets, so that the overall results can show a general trend.

| Causal Data | Learning Task | Sup. BLEU | ΔSSL (BLEU) |
|---|---|---|---|
| En→Cipher | Causal | 19.20 | +1.84 |
| | Anticausal | 7.75 | +38.02 |
| Cipher→En | Causal | 17.08 | +4.05 |
| | Anticausal | 7.97 | +38.01 |

Table 4: SSL improvements (ΔSSL) in BLEU score across causal vs. anticausal learning tasks on the synthetic decipherment datasets.

convenience), (ii) apply the ROT13 encryption algorithm (Schneier, 1996) to obtain the encrypted corpus, and then (iii) apply noise on the corpus that is chosen to be the effect corpus.

In the encryption step (ii), for each English sentence $\boldsymbol{x}$, its encryption $\mathrm{ROT13}(\boldsymbol{x})$ replaces each letter with the 13th letter after it in the alphabet, e.g., "A"→"N," "B"→"O." Note that we choose ROT13 due to its invertibility, since $\mathrm{ROT13}(\mathrm{ROT13}(\boldsymbol{x})) = \boldsymbol{x}$. Therefore, without any noises, the corpus of English and the corpus of encrypted sequences by ROT13 are symmetric.

In the noising step (iii), we apply noise either to the English text or to the ciphertext, thus creating two datasets Cipher→En, and En→Cipher, respectively. When applying noise to a sequence, we use the implementation of the Fairseq library.[11] Namely, we mask some random words in the sequence (word masking), permute a part of the sequence (permuted noise), randomly shift the endings of the sequence to the beginning (rolling noise), and insert some random characters or masks to the sequence (insertion noise). We set the probability of all noises to $p = 5\%$.

**Results** For each of the two datasets En→Cipher and Cipher→En, we perform SSL in the causal and anticausal direction by either treating the input $X$ as the cause and the target $Y$ as the effect, or vice versa. Specifically, we use a standard Transformer architecture for the supervised model, and for SSL, we multitask the translation task with an additional denoising autoencoder (Vincent et al., 2008) using the Fairseq Python package. The results are shown in Tab. 4. It can be seen that in both cases, anticausal models show a substantially larger SSL improvement than causal models.

We also note that there is a substantial gap in the supervised performance between causal and anticausal learning tasks on the same underlying data. This is also expected as causal learning is typically easier than anticausal learning since it corresponds to learning the "natural" forward function, or causal mechanism, while anticausal learning corresponds to learning the less natural, non-causal inverse mechanism.

| Task Type | Mean ΔSSL (±std) | According to ICM |
|---|---|---|
| Causal | +0.04 (±4.23) | Smaller or none |
| Anticausal | +1.70 (±2.05) | Larger |

Table 5: Meta-study of SSL improvement (ΔSSL) across 55 causal and 50 anticausal NLP tasks.

### 5.2 SSL Improvements in Existing Work

After verifying the different behaviour in SSL improvement predicted by the ICM principle on the decipherment experiment, we conduct an extensive meta-study to survey whether this trend is also reflected in published NLP findings. To this end, we consider a diverse set of tasks, and SSL methods. The tasks covered in our meta-study include machine translation, summarization, parsing, tagging, information extraction, review sentiment classification, text category classification, word sense disambiguation, and chunking. The SSL methods include self-training, co-training (Blum and Mitchell, 1998), tri-training (Zhou and Li, 2005), transductive support vector machines (Joachims, 1999), expectation maximization (Nigam et al., 2006), multitasking with language modeling (Dai and Le, 2015), multitasking with sentence reordering (as used in Zhang and Zong (2016)), and cross-view training (Clark et al., 2018). Further details on our meta study are explained in Appendix A.

We covered 55 instances of causal learning and 50 instances of anticausal learning. A summary of the trends of causal SSL and anticausal SSL are listed in Tab. 5. Echoing with the implications of ICM stated in § 3, for causal learning tasks, the average improvement by SSL is only very small, 0.04%. In contrast, the anticausal SSL improvement is larger, 1.70% on average. We use Welch's t-test (Welch, 1947) to assess whether the difference in mean between the two distributions of SSL improvment (with unequal variance) is significant and obtain a p-value of 0.011.

## 6 DA for Causal vs. Anticausal Models

We also consider a supervised domain adaptation (DA) setting in which the goal is to adapt a model trained on a large labeled data set from a source domain, to a potentially different target domain from which we only have a a small labeled data set. As explained in § 3, DA should only work well for causal learning, but not necessarily for anticausal learning, according to the ICM principle.

[11] Link to the Fairseq implementation.

| Task Type | Mean $\Delta$DA ($\pm$std) | According to ICM |
|---|---|---|
| Causal | 5.18 ($\pm$6.57) | Larger |
| Anticausal | 1.26 ($\pm$1.79) | Smaller |

Table 6: Meta-study of DA improvement ($\Delta$DA) across 22 causal and 11 anticausal NLP tasks.

Similar to the meta-study on SSL, we also review existing NLP literature on DA. We focus on DA improvement, i.e., the performance gain of using DA over an unadapted baseline that only learns from the source data and is tested on the target domain. Since the number of studies on DA that we can find is smaller than for SSL, we cover 22 instances of DA on causal tasks, and 11 instances of DA on anticausal tasks.

The results are summarised in Tab. 6. We find that the observations again echo with our expectations (according to ICM) that DA should work better for causal, than for anticausal learning tasks. Again, we use Welch's t-test (Welch, 1947) to verify that the DA improvements of causal learning and anticausal learning are statistically different, and obtain a p-value of 0.023.

## 7 How to Use the Findings in this Study

**Data Collection Practice in NLP** Due to the different implications of causal and anticausal learning tasks, *we strongly suggest annotating the causal direction when collecting new NLP data.* One way to do this is to only collect data from one causal direction and to mention this in the meta information. For example, summarization data collected from the TL;DR of scientific papers SciTldr (Cachola et al., 2020) should be *causal*, as the TL;DR summaries on OpenReview (some from authors when submitting the paper, others derived from the beginning of peer reviews) were likely composed after the original papers or reviews were written. Alternatively, one may allow mixed corpora, but label the causal direction for each $(\boldsymbol{x}, \boldsymbol{y})$ pair, e.g., which is the original vs. translated text in a translation pair. Since more data often leads to better model performance, it is common to mix data from both causal directions, e.g., training on both En→Es and Es→En data. Annotating the causal direction for each pair allows future users of the dataset to potentially handle the causal and anticausal parts of the data differently.

**Causality-Aware Modeling** When building NLP models, the causal direction provides additional information that can potentially be built into the model. In the MT case, since causal and anticausal learning can lead to different performance (Ni et al., 2021), one way to take advantage of the known causal direction is to add a prefix such as "[Modeling-Effect-to-Cause]" to the original input, so that the model can learn from causally-annotated input-output pairs. For example, Riley et al. (2020) use labels of the causal direction to elicit different behavior at inference time. Another option is to carefully design a combination of different modeling techniques, such as limiting self-training (a method for SSL) only to the anticausal direction and allowing back-translation in both directions, as preliminarily explored by Shen et al. (2021).

**Causal Discovery** Suppose that we are given measurements of two types of NLP data $X$ and $Y$ (e.g., text, parse tree, intent type) whose collection process is unknown, i.e., which is the cause and which the effect. One key finding of our study is that there is typically a causal footprint of the data collection process which manifests itself, e.g., when computing the description length in different directions (§ 4) or when performing SSL (§ 5) or DA (§ 6). Based on which direction has the shorter MDL, or allows better SSL or DA, we can thus infer one causal direction over the other.

**Prediction of SSL and DA Effectiveness** Being able to predict the effectiveness of SSL or DA for a given NLP task can be very useful, e.g., to set the weights in an ensemble of different models (Søgaard, 2013). While predicting SSL performance has previously been studied from a non-causal perspective (Nigam and Ghani, 2000; Asch and Daelemans, 2016), our findings suggest that a simple qualitative description of the data collection process in terms of its causal direction (as summarised for the most common NLP tasks in Tab. 1) can also be surprisingly effective to evaluate whether SSL or DA should be expected to work well.

## 8 Limitations and Future Work

We note that ICM—when taken strictly—is an idealized assumption that may be violated and thus may not hold exactly for a given real-world data set, e.g., due to confounding, i.e., when both variables are influenced by a third, unobserved variable. In this case, one may observe less of a difference between causal and anticausal learning tasks.

We also note that, while we have made an effort to classify different NLP tasks as *typically* causal or anticausal, our categorization should not be applied blindly without regard for the specific generative process at hand: deviations are possible as explained in the Mixed/Other category.

Another limitation is that the SSL and DA settings considered in this paper are only a subset of the various settings that exist in NLP. Our study does not cover, for example, SSL that uses additional output data (e.g., Jean et al. (2015); Gülçehre et al. (2015); Sennrich and Zhang (2019)), or unsupervised DA (as reviewed by Ramponi and Plank (2020)). In addition, in our meta-study of published SSL and DA findings, the improvements of causal vs. anticausal learning might be amplified by the scale of research efforts on different tasks and potentially suffer from selection bias.

Finally, we remark that, in the present work, we have focused on bivariate prediction tasks with an input $X$ and output $Y$. Future work may also apply ICM-based reasoning to more complex NLP settings, for example, by (i) incorporating additional (sequential/temporal) structure of the data (e.g., for MT or language modeling) or (ii) considering settings in which the input $X$ consists of both cause $X_{\text{CAU}}$ and effect $X_{\text{EFF}}$ features of the target $Y$ (von Kügelgen et al., 2019, 2020).

## 9 Related Work

**NLP and Causality** Existing work on NLP and causality mainly focuses on the extracting text features for causal inference. Researchers first propose a causal graph based on domain knowledge, and then use text features to represent some elements in the causal graph, e.g., the cause (Egami et al., 2018), effect (Fong and Grimmer, 2016), and confounders (Roberts et al., 2020; Veitch et al., 2020; Keith et al., 2020). Another line of work mines causal relations among events from textual expressions, and uses them to perform relation extraction (Do et al., 2011; Mirza and Tonelli, 2014; Dunietz et al., 2017; Hosseini et al., 2021), question answering (Oh et al., 2016), or commonsense reasoning (Sap et al., 2019; Bosselut et al., 2019). For a recent survey, we refer to Feder et al. (2021).

**Usage of MDL in NLP** Although MDL has been used for causal discovery for low-dimensional data (Budhathoki and Vreeken, 2017; Mian et al., 2021; Marx and Vreeken, 2021), only very few studies adopt MDL on high-dimensional NLP data. Most existing uses of MDL on NLP are for probing and interpretability: e.g., Voita and Titov (2020) use it for probing of a small Bayesian model and network pruning, based on the method proposed by Blier and Ollivier (2018) to calculate MDL for deep learning. We are not aware of existing work using MDL for causal discovery, or to verify causal concepts such as ICM in the context of NLP.

**Existing Discussions on SSL and DA in NLP** SSL and DA has long been used in NLP, as reviewed by Søgaard (2013) and Ramponi and Plank (2020). However, there have been a number of studies that report negative results for SSL (Clark et al., 2003; Steedman et al., 2003; Reichart and Rappoport, 2007; Abney, 2007; Spreyer and Kuhn, 2009; Søgaard and Rishøj, 2010) and DA (Plank et al., 2014). Our works constitutes the first explanation of the ineffectiveness of SSL and DA on certain NLP tasks from the perspective of causal and anticausal learning.

## 10 Conclusion

This work presents the first effort to use causal concepts such as the ICM principle and the distinction between causal and anticausal learning to shed light on some commonly observed trends in NLP. Specifically, we provide an explanation of observed differences in SSL (Tabs. 4 and 5) and DA (Tab. 6) performance on a number of NLP tasks: DA tends to work better for causal learning tasks, whereas SSL typically only works for anticausal learning tasks, as predicted by the ICM principle. These insights, together with our categorization of common NLP tasks (Tab. 1) into causal and anticausal learning, may prove useful for future NLP efforts. Moreover, we empirically confirm using MDL that the description of data is typically shorter in the causal than in the anticausal direction (Tab. 3), suggesting that a causal footprint can also be observed for text data. This has interesting potential implications for discovering causal relations between different types of NLP data.

## Acknowledgements

We thank Simon Buchholz for helpful discussions, Nasim Rahaman, Shehzaad Dhuliawala, Yifan Hou, Tiago Pimentel, and the anonymous reviewers for feedback on the manuscript, and Di Jin for helping with computational resources. This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B, and by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645.

## Ethical Considerations

**Use of Data** This paper uses two types of data, a subset of existing machine translation dataset, and synthetic decipherment data. As far as we are concerned, there are no sensitive issues such as privacy regarding the data usage.

**Potential Stakeholders** This research focuses on meta properties of two commonly applied methodologies, SSL and DA in NLP. Although this research is not directly connected to specific applications in society, the usage of this study can benefit future research in SSL and DA.

## References

Steven Abney. 2007. Semisupervised learning for computational linguistics.

Vincent Van Asch and Walter Daelemans. 2016. Predicting the effectiveness of self-training: Application to sentiment classification. *CoRR*, abs/1601.03288.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Léonard Blier and Yann Ollivier. 2018. The description length of deep learning models. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Avrim Blum and Tom M. Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory, COLT 1998, Madison, Wisconsin, USA, July 24-26, 1998*, pages 92–100. ACM.

Nikolay Bogoychev and Rico Sennrich. 2019. Domain, translationese and noise in synthetic data for neural machine translation. *CoRR*, abs/1911.03362.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Kailash Budhathoki and Jilles Vreeken. 2017. MDL for causal inference on discrete data. In *2017 IEEE International Conference on Data Mining, ICDM 2017, New Orleans, LA, USA, November 18-21, 2017*, pages 751–756. IEEE Computer Society.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734. ACL.

Kevin Clark, Minh-Thang Luong, Christopher D Manning, and Quoc V Le. 2018. Semi-supervised sequence modeling with cross-view training. *arXiv preprint arXiv:1809.08370*.

Stephen Clark, James Curran, and Miles Osborne. 2003. Bootstrapping POS-taggers using unlabelled data. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 49–55.

Andrew M. Dai and Quoc V. Le. 2015. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3079–3087.

P. Daniušis, D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. 2010. Inferring deterministic causal relations. In *26th Conference on Uncertainty in Artificial Intelligence*, pages 143–150, Corvallis, OR. AUAI Press. Best student paper award.

Quang Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017. The BECauSE corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104, Valencia, Spain. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846, Online. Association for Computational Linguistics.

Naoki Egami, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2018. How to make causal inferences using texts. *CoRR*, abs/1802.02163.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brando n M. Stewart, Victor Veitch, and Diyi Yang. 2021. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *CoRR*, abs/2109.00725.

Christian Fong and Justin Grimmer. 2016. Discovery of treatments from text corpora. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. 2016. Domain adaptation with conditional transferable components. In *International conference on machine learning*, pages 2839–2848. PMLR.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Peter D Grünwald. 2007. *The minimum description length principle*. MIT press.

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *CoRR*, abs/1503.03535.

Nizar Habash. 2008. Four techniques for online handling of out-of-vocabulary words in arabic-english statistical machine translation. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA, Short Papers*, pages 57–60. The Association for Computer Linguistics.

Pedram Hosseini, David A. Broniatowski, and Mona T. Diab. 2021. Predicting directionality in causal relations in text. *CoRR*, abs/2103.13606.

Hal Daumé III and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA - Short Papers*, pages 407–412. The Association for Computer Linguistics.

Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. 2012. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31.

Dominik Janzing and Bernhard Schölkopf. 2010. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194.

Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. Montreal neural machine translation systems for wmt'15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 134–140.

Thorsten Joachims. 1999. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999), Bled, Slovenia, June 27 - 30, 1999*, pages 200–209. Morgan Kaufmann.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Katherine A. Keith, David Jensen, and Brendan O'Connor. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5332–5344. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Andrei N Kolmogorov. 1965. Three approaches to the quantitative definition of information. *Problems of information transmission*, 1(1):1–7.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Jan Lemeire and Erik Dirkx. 2006. Causal models as minimal descriptions of multivariate systems.

Ming Li, Paul Vitányi, et al. 2008. *An introduction to Kolmogorov complexity and its applications*, volume 3. Springer.

Alexander Marx and Jilles Vreeken. 2021. Formally justifying mdl-based inference of cause and effect. *CoRR*, abs/2105.01902.

Osman Mian, Alexander Marx, and Jilles Vreeken. 2021. Discovering fully oriented causal networks. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Paramita Mirza and Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2097–2106, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Jingwei Ni, Zhijing Jin, Markus Freitag, Mrinmaya Sachan, and Bernhard Schoelkopf. 2021. Original or translated? causal effects of data collection direction onmachine translation performance.

Kamal Nigam and Rayid Ghani. 2000. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the 2000 ACM CIKM International Conference on Information and Knowledge Management, McLean, VA, USA, November 6-11, 2000*, pages 86–93. ACM.

Kamal Nigam, Andrew McCallum, and Tom M. Mitchell. 2006. Semi-supervised text classification using EM. In Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors, *Semi-Supervised Learning*, pages 32–55. The MIT Press.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Ryu Iida, Masahiro Tanaka, and Julien Kloetzer. 2016. A semi-supervised learning approach to why-question answering. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3022–3029. AAAI Press.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Barbara Plank, Anders Johannsen, and Anders Søgaard. 2014. Importance weighting and unsupervised domain adaptation of POS taggers: a negative result. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 968–973. ACL.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Roi Reichart and Ari Rappoport. 2007. Self-training for enhancement and domain adaptation of statistical parsers trained on small datasets. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 616–623, Prague, Czech Republic. Association for Computational Linguistics.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online. Association for Computational Linguistics.

Jorma Rissanen. 1984. Universal coding, information, prediction, and estimation. *IEEE Trans. Inf. Theory*, 30(4):629–636.

Margaret E Roberts, Brandon M Stewart, and Richard A Nielsen. 2020. Adjusting for confounding with text matching. *American Journal of Political Science*, 64(4):887–903.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press.

Bruce Schneier. 1996. Applied cryptography," john willey & sons. *Inc.*,.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris M. Mooij. 2012. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Eleni Sgouritsa, Dominik Janzing, Philipp Hennig, and Bernhard Schölkopf. 2015. Inference of cause and effect with unsupervised inverse regression. In *Artificial intelligence and statistics*, pages 847–855. PMLR.

Jiajun Shen, Peng-Jen Chen, Matthew Le, Junxian He, Jiatao Gu, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2021. The source-target domain mismatch problem in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1519–1533, Online. Association for Computational Linguistics.

Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244.

Anders Søgaard. 2013. Semi-supervised learning and domain adaptation in natural language processing. *Synthesis Lectures on Human Language Technologies*, 6(2):1–103.

Anders Søgaard and Christian Rishøj. 2010. Semi-supervised dependency parsing using generalized tri-training. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1065–1073, Beijing, China. Coling 2010 Organizing Committee.

Ray J Solomonoff. 1964. A formal theory of inductive inference. part ii. *Information and control*, 7(2):224–254.

Kathrin Spreyer and Jonas Kuhn. 2009. Data-driven dependency parsing of new languages using incomplete and noisy training data. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 12–20, Boulder, Colorado. Association for Computational Linguistics.

Mark Steedman, Miles Osborne, Anoop Sarkar, Stephen Clark, Rebecca Hwa, Julia Hockenmaier, Paul Ruhlen, Steven Baker, and Jeremiah Crim. 2003. Bootstrapping statistical parsers from small datasets. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary. Association for Computational Linguistics.

Masashi Sugiyama and Motoaki Kawanabe. 2012. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press.

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus - parallel and free: `http://logos.uio.no/opus`. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Victor Veitch, Dhanya Sridhar, and David M. Blei. 2020. Adapting text embeddings for causal inference. In *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3-6, 2020*, volume 124 of *Proceedings of Machine Learning Research*, pages 919–928. AUAI Press.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 1096–1103. ACM.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 183–196. Association for Computational Linguistics.

Julius von Kügelgen, Alexander Mey, and Marco Loog. 2019. Semi-generative modelling: Covariate-shift adaptation with cause and effect features. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1361–1369. PMLR.

Julius von Kügelgen, Alexander Mey, Marco Loog, and Bernhard Schölkopf. 2020. Semi-supervised learning, causality, and the conditional cluster assumption. In *Conference on Uncertainty in Artificial Intelligence*, pages 1–10. PMLR.

Bernard L Welch. 1947. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.

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. Association for Computational Linguistics.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.

Kun Zhang, Mingming Gong, and Bernhard Schölkopf. 2015. Multi-source domain adaptation: A causal view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. 2013. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827. PMLR.

Zhi-Hua Zhou and Ming Li. 2005. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Trans. Knowl. Data Eng.*, 17(11):1529–1541.

## A Meta Study Settings of SSL and DA

For the meta study of SSL, we covered but are not limited to all relevant papers cited by the review on NLP SSL by Søgaard (2013). We went through the leaderboard of many NLP tasks and covered the SSL papers listed on the leaderboards. The papers covered by our meta study are available on our GitHub.

For supervised DA, we searched papers with the keyword domain adaptation and task names from a wide range of tasks that use supervised DA.

Note that for fair comparison, we do not consider papers without a comparable supervised baseline corresponding to the SSL, or a comparable unadapted baseline corresponding to the DA. We do not consider MT DA which tackles the out-of-vocabulary (OOV) problem because $P(E|C)$ may be different for OOV (Habash, 2008; III and Jagarlamudi, 2011).

## B Experimental Details of Minimum Description Length

We calculate the MDL(X) and MDL(Y) by a language model, and obtain MDL(X|Y) and MDL(Y|X) using translation models. For language model, we use the autoregressive GPT2 (Radford et al., 2019), and for the translation model, we the Marian Neural Machine Translation model (Junczys-Dowmunt et al., 2018) trained on the OPUS Corpus (Tiedemann and Nygaard, 2004). Both these models use the layers from the transformer model (Vaswani et al., 2017). The autoregressive language model consists only of decoder layers, whereas the translation model used six encoder and six decoder layers. Both of these models have roughly the same number of parameters. We used the huggingface implementation (Wolf et al., 2020) of these models for their respective set of languages.