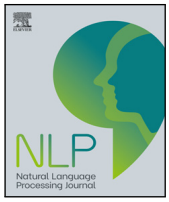
Natural Language Processing Journal 6 (2024) 100055

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp



On the relation between K–L divergence and transfer learning performance on causality extraction tasks

Seethalakshmi Gopalakrishnan [a,*], Victor Zitian Chen [b], Wenwen Dou [a], Wlodek Zadrozny [a]

[a] University of North Carolina at Charlotte, 9201 University City Blvd, Charlotte, NC 28223, USA
[b] Fidelity Investments, 100 New Millennium Way, Durham, NC 27709, USA

ARTICLE INFO

Keywords:
Transfer learning
Domain adaptability
Causality extraction
Large language models
Kullback–Leibler divergence
BERT
DistilBERT
NLP
Natural language processing

ABSTRACT

The problem of extracting causal relations from text remains a challenging task, even in the age of Large Language Models (LLMs). A key factor that impedes the progress of this research is the availability of the annotated data and the lack of common labeling methods. We investigate the applicability of transfer learning (domain adaptation) to address these impediments in experiments with three publicly available datasets: FinCausal, SCITE, and Organizational. We perform pairwise transfer experiments between the datasets using DistilBERT, BERT, and SpanBERT (variants of BERT) and measure the performance of the resulting models. To understand the relationship between datasets and performance, we measure the differences between vocabulary distributions in the datasets using four methods: Kullback–Leibler (K–L) divergence, Wasserstein metric, Maximum Mean Discrepancy, and Kolmogorov–Smirnov test. We also estimate the predictive capability of each method using linear regression. We record the predictive values of each measure. Our results show that K–L divergence between the distribution of the vocabularies in the data predicts the performance of the transfer learning with R2 = 0.0746. Surprisingly, the Wasserstein distance predictive value is low (R2=0.52912), and the same for the Kolmogorov–Smirnov test (R2 =0.40025979). This is confirmed in a series of experiments. For example, with variants of BERT, we observe an almost a 29% to 32% increase in the macro-average F1-score, when the gap between the training and test distributions is small, according to the K–L divergence — the best-performing predictor on this task. We also discuss these results in the context of the sub-par performance of some large language models on causality extraction tasks. Finally, we report the results of transfer learning informed by K–L divergence; namely, we show that there is a 12 to 63% increase in the performance when a small portion of the test data is added to the training data. This shows that corpus expansion and n-shot learning benefit, when the process of choosing examples maximizes their information content, according to the K–L divergence.

1. Introduction

Causality extraction is one of the challenging tasks in information extraction. It is the process of extracting cause-and-effect relationships from the text. For example, consider the following sentence from the 2020 SEC 10-K Documents of 65 S&P 80 Financial Companies:

When a policyholder or insured gets sick or hurt, the company pays cash benefits fairly and promptly for eligible claims.

There exists a causal relationship between the cause "*policyholder or insured gets sick or hurt*" and the effect "*cash benefits fairly and promptly for eligible claims*".

The process of extracting the cause–effect relationships from the text is called *causality extraction*. Such relations can be present in texts from various domains and, if extracted, can be used for applications including question answering (Hassanzadeh et al., 2019; Dang, 2021; Girju, 2003; Sobrino et al., 2014), information retrieval (Khoo et al., 2001), and medical text mining (Ding et al., 2019a). Despite the importance of the causality extraction problem and increased attention on this task in recent times, data insufficiency remains a challenge and an open research problem (Yang et al., 2022).

Transfer learning, or domain adaptation, (Bommasani et al., 2021; Pan and Yang, 2009) have been proposed as a mitigation for the problem of scarcity of annotated data. The idea is that the performance of a machine learning program can be enhanced by pretraining on a related task. A survey article (Weiss et al., 2016) formally defines it as follows:

> "Given a source domain $\mathtt{D}_S$ with a corresponding task $\mathtt{T}_S$ and a target domain $\mathtt{D}_T$ with corresponding task $\mathtt{T}_T$, where $\mathtt{D}_S \neq \mathtt{D}_T$ or

* Corresponding author.
E-mail addresses: sgopala4@uncc.edu (S. Gopalakrishnan), founder@gopeaks.org (V.Z. Chen), wdou1@uncc.edu (W. Dou), wzadrozn@uncc.edu (W. Zadrozny).

https://doi.org/10.1016/j.nlp.2024.100055
Received 4 August 2023; Received in revised form 5 January 2024; Accepted 11 January 2024

$T_S \neq T_T$, transfer learning aims to improve the performance of the model's predictions by using the related information from D_S and T_S."

This definition obviously raises the question how we should measure the difference between domains or tasks. In this article, we use the Kullback–Leibler divergence (K–L divergence) (Manning, 2009) to measure the differences between distributions of terms in the datasets. We also report results using the Wasserstein distance and the Kolmogorov–Smirnov test (Gibbs and Su, 2002).

For our experiments, we use three datasets Organizational dataset (ORG) (Gopalakrishnan et al., 2023) that is annotated on financial text, SCITE (Li et al., 2021b), which is annotated on texts from the web, and FinCausal (Mariko et al., 2022), which is annotated on the financial news articles.

In the experiments, for causality extraction, we use DistilBERT (Sanh et al., 2019), https://huggingface.co/docs/transformers/model_doc/distilbert, a variant of BERT (Devlin et al., 2019) (still, one state-of-the-art performing models), we also run the same set of experiments with BERT and SpanBERT.

Our experiments show that different distributions of word frequencies in the datasets lead to different success rate in transfer learning (which we discuss in Section 5).

We show that K–L divergence can be a basis for improving transfer learning via corpus expansion/n-shot learning. When the value of the K–L divergence is reduced by adding a small portion of the test data to the train data. We show that there is a performance improvement from 12% to 63% in all sets of experiments. When we add domain-specific data, such as financial text to the dataset that is created on a data from web search such as SCITE, the improvement is higher.

2. Preliminaries and related work

An objective of this work is to understand the potential of transfer learning to improve the accuracy of causality extraction from text. To be more precise, we want to understand the relationship between the properties of datasets and the degree of success in transfer learning. Therefore in this preliminaries we provide some pointers that help explain our methods and put the results in context of prior work.

We first discuss measures of difference between datasets, then provide a few references to transfer learning (sometimes called 'domain adaptation'), and finally to selected prior work on causality extraction.

2.1. Measures of divergence and their uses

There are infinitely many ways we can talk about differences between text data. However, the simplest measures of difference count-based, i.e. statistical. We use three popular tests for differences between the distributions: Kullback–Leibler divergence, Wasserstein distance, and Kolmogorov–Smirnov test. The mathematical relations between them are described in Gibbs and Su (2002). However, in this article we care about their potential predictive powers with respect to the accuracy of transfer learning for causality extraction (see Section 5). All three tests have been used in NLP, and K–L divergence is perhaps the most popular.

K–L divergence is a statistical distance that measures how different is a probability distribution compared to another. It is denoted by $D_{KL}(P \mid\mid Q)$ where P and Q are the probability distributions (Manning, 2009). Notably, it is not symmetric $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$.

An example recent use is shown in Li et al. (2021a) to disentangle the syntax and semantics in a deep decomposable model. For semantic similarity tasks and syntactic similarity tasks, their model improves their disentanglement quality.

In this article, using K–L divergence we show that the distributions of the three datasets are different and predict quite well transfer results.

K–L Divergence measures the difference between two probability distributions based on information theory, that is how much we can learn from one distribution about another, and therefore is not symmetric. The other two measures are symmetric. Wasserstein distance measures the distance between two probability distributions by considering the 'cost' of transforming one into the other and is symmetric. That is why it is sometimes called "the earth mover distance". Finally, the Kolmogorov–Smirnov test is a statistical test used to compare empirical distributions and is often employed to determine if a sample comes from a specific distribution and is not symmetric.

In this article we are applying these concepts without any modifications. The reader can find an introduction, formulas and comparisons of mathematical properties in Gibbs and Su (2002). Examples of their uses in NLP appear e.g. in Manning (2009), Martin (2009), Chen et al. (2018) and Al Kuwatly et al. (2020). For our practical objectives, we care about the existence of (Python) packages that we can easily apply to compute the required measures. We will provide references to them in Section 4.

2.2. Transfer learning

Many machine-learning models perform well under the assumption that the train and the test data have the same distribution and feature space (Pan and Yang, 2010). If the distribution differs, the model has to be built from scratch by annotating a new dataset for that particular domain. This process of annotating a new dataset for every domain will be a challenging process and expensive one. Transfer learning between the task domains should be helpful in such scenarios. However, as shown in Zoph et al. (2020), transfer does not always produce positive results.

In NLP, transfer has been used in the Natural Language Inference (NLI) task (e.g. Conneau et al. (2017)), and for various other tasks like causal sentence detection (Kyriakakis et al., 2019); finding condition-action sentences in medical guidelines (Hematialam and Zadrozny, 2021) and understanding of biomedical texts (Peng et al., 2019). In another example, extracting drug timelines from the electronic health record was done by Miller et al. (2021) by training the model on THYME colon cancer corpus and testing on THYME brain cancer corpus.

We need to note that the term *transfer learning* is often used interchangeably the term *domain adaptation*, especially in natural language processing, as observed by Pan and Yang (2009). In Sun et al. (2015) we see the following definition "*domain adaptation* is a subcategory of transfer learning. In domain adaptation, the source and target domains all have the same feature space (but different distributions); in contrast, *transfer learning* includes cases where the target domain's feature space is different from the source feature space or spaces."

In our case, the feature spaces (the inputs to the variants of BERT) are different because of the differences in the vocabularies, and also we have the differences in feature distributions (Section 3). We felt that perhaps 'transfer learning' is a better fitting term, but we also added 'domain adaptation' parenthetically in the introduction.

In general, transfer learning and domain adaptation are very active areas of research in machine learning, necessitating dozens of survey articles every year.

2.3. Causality extraction

BERT (Devlin et al., 2018) has been used both to test transfer learning and domain adaptability for data extraction from text. As a result many variants of BERT have been created, including DistilBERT (Sanh et al., 2019), SpanBERT (Joshi et al., 2020) used in this article.

Even with the appearance of larger language models, it can give a state-of-the-art performance for causality extraction tasks (Khetan et al., 2020; Lyu et al., 2022; Gopalakrishnan et al., 2023; Peng et al., 2019). Similarly, it performs well on other tasks involving domain adaptability such as sentiment classification (Rietzler et al., 2019),

hate speech detection (Mozafari et al., 2020), Biomedical Named Entity Recognition (Sun and Yang, 2019).

A recent survey on causality extraction (Yang et al., 2022) classifies the existing methods for causality extraction into knowledge-based, statistical-machine-learning-based, and deep-learning-based methodologies. Initial works on causality extraction used rules and linguistic features (Garcia et al., 1997; Radinsky et al., 2012; Kang et al., 2014; Bui et al., 2010). Statistical-machine-learning-based models can use linguistic features, verb-pair rules, etc., as well as discourse features, to train the classifiers such as Naive Bayes and Support Vector Machines, etc. Gu et al. (2016) and Pakray and Gelbukh (2014). Recently, deep learning-based models have been used for the causality extraction task (Zhang et al., 2018; Li et al., 2017).

Other related work in this space include (Peng et al., 2021; Li, 2022). The first article compares the performance of BERT and FinBERT for the financial text processing tasks, and studies how different types of pre-training affects the system's performance. The second article studies performance of the information extraction models on the complex conversations using different domains, and proposes a causality method to learn the distribution shifts in the data, and uses causal inference frameworks to reason about these shifts.

Causality extraction methods can also be used for finding other information of interest, such as emotions and their causes. Xia and Ding (2019) aim to extract the emotion–cause pairs by annotating a corpus for emotion–cause pair extraction. First, they extract the emotions and the causes in the text individually, and then create the cause–emotion pairs and filter them using a Bi-LSTM. Similar work, presented in Ding et al. (2019b), aims to identify potential causes that lead to emotions using a Bi-LSTM and attention, using a corpus consisting of texts, their relative positions, and global labels that record the predictions of the previous clause to record the predictions of the previous clauses. A more recent work (Chen et al., 2022) addresses the problem of cascading errors (incorrect information propagated in a pipeline model) by introducing reinforcement learning. It utilizes the BERT semantic embeddings and a Bi-LSTM for emotion–cause extraction. Chen et al. (2023) aim to determine the causal relationships between the input pair of emotion and cause. They also show how to extracts specific context clause in causal relationships using a combination of an Albert transformer-based model and a Bi-LSTM.

Causality extraction is a rapidly growing sub-field of NLP, and the above presents only a sample of existing approaches to the problem. No doubt, new methods will be developed based on newer large language models and increasing amounts of annotated data.

3. Data

In our experiments, we use three causality extraction datasets. First, SCITE (Li et al., 2021b), which extends the annotations of SemEval 2010 task 8 dataset (Hendrickx et al., 2019) by considering all the causal triplets present in the sentence, whereas (Hendrickx et al., 2019) considers only one causal triplet in the sentence. This dataset consists of text data from the web, which is not particularly related to the financial domain. Second, FinCausal (Mariko et al., 2022), which is created as part of a challenge FinCausal 2022. This challenge aims to extract causalities from financial documents. This data is extracted from the 2019 financial news, which is collected from 14,000 economics and finance websites. Third, the Organizational (ORG) dataset, which was created for the causality extraction on Financial documents (Gopalakrishnan et al., 2023). In this dataset, the 2020 SEC 10-K Documents of 65 S&P 80 Financial Companies were collected and manually annotated.

Here are some examples from each dataset. In the SCITE dataset, the cause–effect pairs are annotated using the XML tags, as shown below:

Example 1 — SCITE

```
<item id="15" label="Cause-Effect((e1,e2))"> <sente
nce> This case arises from <e1>;a December 21, 2005
automobile accident <e1> that resulted in <e2> the
death <e2> of Larry Haynes.</sentence>
```

In the FinCausal dataset, the cause–effect relation pairs are available as tags. <e1> represents the cause and <e2> represents the effect. The phrases of cause/effect is also available.

Example 2 — FinCausal

```
Text: It found that total U.S. healthcare spending
would be about $3.9 trillion under Medicare for All
in 2019, compared with about $3.8 trillion under the
status quo. Part of the reason is that Medicare for
All would offer generous benefits with no copays and
deductibles, except limited cost-sharing for certain
medications.
```

```
Tag format: <e2>It found that total U.S. healthcare
spending would be about $3.9 trillion under Medicare
for All in 2019, compared with about $3.8 trillion un-
der the status quo.</e2> <e1>Part of the reason is that
Medicare for All would offer generous benefits with no
copays and deductibles, except limited cost-sharing
for certain medications.</e1>
```

The organizational data is annotated in the BIO-label format. For each of the tokens in the text, a label will be assigned. The cause is represented as C, effect as E.

Example 3 — Organizational

```
text:      ["When","a","policyholder","or","insured",
"gets","sick","or","hurt","the","Company","pays",
"cash","benefits","fairly","and","promptly","for", "el-
igible","claims"]
Label: ["O","O","B-C","I-C","I-C","I-C","I-C","I-C", "I-
C",    "O",    "O","B-CT","B-E","I-E","I-E","I-E","I-E",
"I-E","I-E", "I-E"]
```

For the transfer learning experiments, we converted all these 3 datasets into the IO label format, like the one above.

4. Methods: Data analysis and the model

Since it is an objective of this article to investigate how the differences in text data impact the performance of a causality extraction model on a new dataset, we first quantify these differences. Then we briefly describe the models used — fine-tuned versions of DistilBERT (Sanh et al., 2019), https://huggingface.co/docs/transformers/model_doc/distilbert, BERT-base-cased (Devlin et al., 2018) and SpanBERT (Joshi et al., 2020).

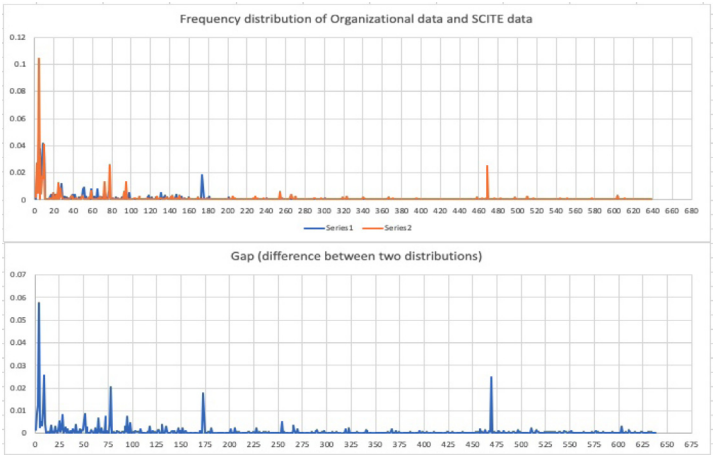
4.1. Differences between the datasets

To understand the differences in the distributions, we have created a feature distribution chart. This chart plots the frequency of the words in both the training and test data.

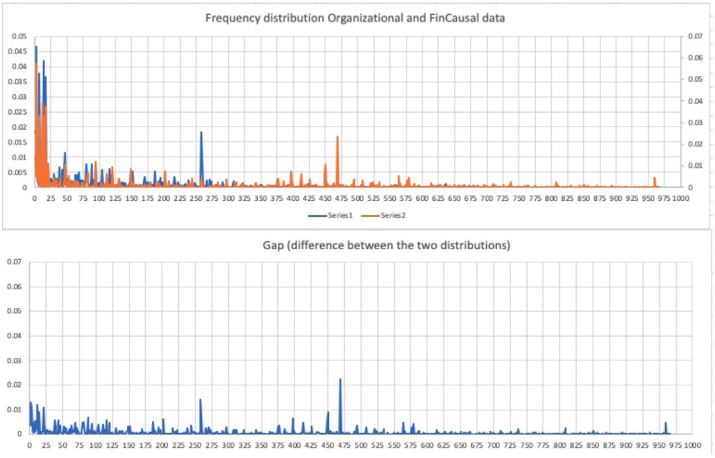
The Organizational data (training data) had a total word count of 4747, and the SCITE data (test data) had a word of 1488. Totally 638 words were common in both of these datasets. Similarly, we have computed a frequency distribution chart for the Organizational and FinCausal dataset. In the FinCausal data, we had a total of 1595 words, out of which 966 are common in both datasets.

These differences in word distributions are shown in Fig. 1 (and further quantified in Table 1). Looking into Figs. 1(a) and 1(b), we can see the gaps between the pairs frequency distributions. The gap between FinCausal and Organizational data seems smaller that for the SCITE data.

This is intuitively explained by the fact that the Organizational and the SCITE data are from completely different domains. Organizational data is created on the financial documents, whereas the SCITE is from the web text. Organizational and FinCausal seem to be similar data because they both are created using the financial text. But the



(a) The top part of the chart with series1 and series2 indicates the frequency distribution on a 638 common word count between the Organizational and SCITE data. The bottom part, with only one blue legend, shows the gap between the two distributions at the top.



(b) The top part of the chart with series1 and series2 indicates the frequency distribution on a 966 common word count between the Organizational and FinCausal data. The bottom part, with only one blue legend, shows the gap between the two distributions at the top.

Fig. 1. From the top and the bottom panel we observe that the difference between the distributions are high for Organizational and SCITE, whereas the gap between the Organizational and FinCausal is smaller. As we shall see later, the difference in distribution is predictive of the F1-score in transfer learning. This is true both, when we measure the differences by the K–L divergence and by the Wasserstein distance, although the former is more accurate. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Summary of the computed K–L divergence values, Wasserstein distance, Kolmogorov–Smirnov test, and Maximum Mean Discrepancy (MMD) on all the combinations of the datasets. The K–L divergence, MMD, and the Wasserstein distance on the same datasets is zero, meaning that there is a maximum overlap between the train and the test datasets. The higher the value of the K–L divergence, the lower the similarity between the datasets. For the K-S test, the low p-values prove that the distributions are different. From the computed values, we can understand that SCITE is less similar to FinCausal, and Organizational datasets.

Train data	Test data	K–L divergence	Wasserstein distance	Kolmogorov–Smirnov	MMD
SCITE	SCITE	0	0	1.0	0
	FinCausal	0.942	4.95	3.9555e−59	0.7796
	Organizational	0.906	13.1	1.1294e−136	0.7982
FinCausal	FinCausal	0	0	1.0	0
	SCITE	0.771	4.95	0.000109	0.1067
	Organizational	0.286	8.15	2.2483e−66	0.1488
Organizational	Organizational	0	0	1.0	0
	FinCausal	0.279	8.15	1.1639e−65	0.1481
	SCITE	0.336	13.1	3.3517e−151	0.3559

Organizational data is annotated on the financial company reports, whereas FinCausal data is annotated on the financial text from the web.

As shown in Table 1, the K–L divergences between the datasets vary, and confirm the impressions from Fig. 1. Thus for the Organizational data and SCITE data we get the values 0.336 and 0.369; in contrast, for Organizational and FinCausal we get 0.279 and 0.286. In both cases and all directions the values are relatively high, which means the distributions are different.

We repeated the same set of comparisons using the Wasserstein metric and the Kolmogorov–Smirnov test. The Wasserstein distance between the Organizational data and the SCITE data is 13.09, and the distance between the Organizational and FinCausal data is 8.14. We got a Wasserstein distance of 0 between the dataset with itself, and the distance between SCITE and FinCausal is 4.95.

Kolmogorov–Smirnov (KS) test can be used to compare two probability distributions to check whether they are drawn from the same

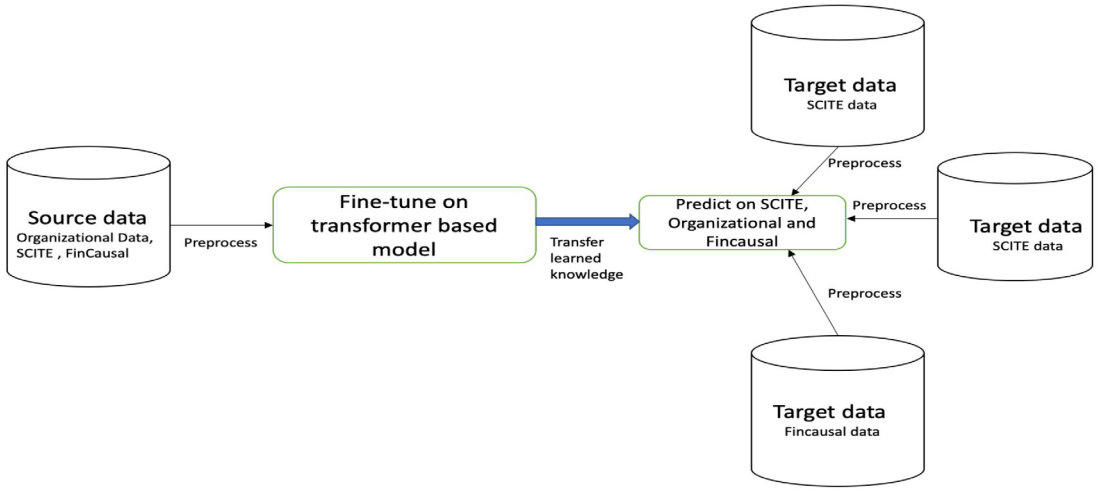


Fig. 2. A schematic description of the experiments. The source/training data are the causal annotated data The knowledge gained during training the model on one of the datasets will be used to predict the other two datasets. For example, training can be done on Organizational and prediction on FinCausal and SCITE data. The data should be preprocessed into IO label format, with special marks for C (Cause) and E (Effect), with the remaining tokens marked as O.

distribution. We chose the standard confidence level of 95%, which means the values that are in favor of the alternative will be rejected if the p-value is less than 0.05. All the p-values we obtained were much smaller than that, indicating and quantifying the differences between the word frequency distributions.

For both computation we used the SCIPY packages: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wasserstein_distance.html, https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ks_2samp.html#scipy.stats.ks_2samp

The Maximum Mean Discrepancy (MMD) is a non-parametric statistical test that can be used to identify the amount of discrepancy between the two probability distributions. It can be given as the distance between the feature means. The MMD between the same datasets will be 0. The maximum the MMD value, the maximum discrepancy between the source and target domain. All the obtained values of differences between the datasets are summarized in Table 1. The table also suggests their intended experimental use in transfer learning. That is, training on one set and testing on another. The impact of these measured differences and the predictive value of each test is discussed in Section 5.

4.2. Models

Given the performance of DistilBERT is good for the causality extraction task on all three datasets individually ((Li et al., 2021b; Mariko et al., 2022; Gopalakrishnan et al., 2023)), in this paper, we ask the next natural question, namely, what happens when we attempt transfer learning, and if there are differences in performance what are they due to. Apart from that, we have also tried BERT, which is the base version, and SpanBERT, which is designed to predict spans of texts. Given the performance of SpanBERT is good for FinCausal data (Li et al., 2021b), we have chosen SpanBERT to compare its performance with its variant BERT and DistilBERT.

5. Experiments and results

To answer this question, we ran several transfer learning experiments with the three datasets. In all the experiments, the DistilBERT, BERT, and SpanBERT models were fine-tuned on one of the datasets, and the other two were used as the test data.

We train our model for (the optimal) 3 epochs with a batch size of 16. All the experiments are conducted on NVIDIA-SMI 525 GPU (using Google Colab).

Fig. 2 gives a schematic overview of the transfer learning experiments. For the FinCausal dataset and Organizational data, several BERT variants perform well, as reported in Lyu et al. (2022) and Gopalakrishnan et al. (2023). In another example, experiments using DistilBERT on the SCITE data produce a (relatively good) macro average F1-score of 0.88 (in our experiment). Here, and later, the results are reported using macro-average scores; for example, F1 refers to the macro-average F1 score, i.e. the average F1-score for all the labels.

Before we discuss the results, we need to mention the composition of the datasets. Thus, SCITE contains the gold annotation for all train, validation, and test subsets. In contrasts, in FinCausal, although, we have splits into train, validation and test sets, there is no gold standard released for the test set, and therefore we use the validation set as test data. In Organizational (2235 sentences), we do the split train (70%), validation (10%), and test (20%).

We performed three sets of rounds of transfer experiments. In the first round of experiments, we fine-tuned the DistilBERT, BERT, and SpanBERT model on the SCITE train dataset, and we tested it on SCITE test data and on validation data of both FinCausal and Organizational test data. In the second round, we fine-tuned on the FinCausal train data and tested on the FinCausal validation and on Organizational and SCITE test data. And in the third set of experiments with transfer, we first fine-tuned on the Organizational training data, and then tested on Organizational and SCITE test data and on FinCausal Validation data.

We are reporting our results of fine-tuning DistilBERTon the FinCausal data and predicting on the FinCausal, even though earlier results of Lyu et al. (2022) are available, showing the F1 of 87.31% on the validation data. Our results, for comparison, are 92%, as shown in Table 2. The difference perhaps due to the fact that we use the Trainer() from the Huggingface to fine-tune the model (https://huggingface.co/docs/transformers/tasks/token_classification), whereas (Lyu et al., 2022) use the transformer model from the Huggingface source (git clone https://github.com/huggingface/transformers.git).

The result of fine-tuning DistilBERT on the Organizational data and testing it on the Organizational data were obtained earlier and are presented in Gopalakrishnan et al. (2023).

5.1. Results and their dependence on the K–L divergence

The results of running the causality extraction task on the SCITE, FinCausal, and Organizational dataset using DistilBERT, BERT, and SpanBERT are summarized in Tables 2, 3, and 4 respectively. All the results are the average of 10 runs.

Table 2
Summary of the transfer learning experiments. This table shows the performance of DistilBERT (DistilBERT-base-cased) for causality extraction. The scores are the average of 10 runs.

Train data	Test data	P	R	F1
SCITE	SCITE	0.86	0.72	0.73
	FinCausal	0.38	0.55	0.12
	Organizational	0.40	0.28	0.16
FinCausal	FinCausal	0.91	0.93	0.92
	SCITE	0.48	0.56	0.31
	Organizational	0.71	0.59	0.59
Organizational	Organizational	0.78	0.78	0.78
	FinCausal	0.40	0.66	0.39
	SCITE	0.32	0.40	0.26

Table 3
Summary of the transfer learning experiments. This table shows the performance of BERT (BERT-base-cased) for causality extraction. The scores are the average of 10 runs.

Train data	Test data	P	R	F1
SCITE	SCITE	0.91	0.77	0.79
	FinCausal	0.39	0.54	0.10
	Organizational	0.41	0.28	0.16
FinCausal	FinCausal	0.91	0.93	0.92
	SCITE	0.48	0.56	0.31
	Organizational	0.76	0.62	0.63
Organizational	Organizational	0.32	0.41	0.29
	FinCausal	0.41	0.66	0.39
	SCITE	0.33	0.41	0.27

Table 4
Summary of the transfer learning experiments. This table shows the performance of SpanBERT ((SpanBERT-large-cased)) for causality extraction. The scores are the average of 10 runs.

Train data	Test data	P	R	F1
SCITE	SCITE	0.92	0.91	0.92
	FinCausal	0.41	0.3	0.2
	Organizational	0.41	0.3	0.2
FinCausal	FinCausal	0.93	0.95	0.94
	SCITE	0.48	0.58	0.32
	Organizational	0.77	0.65	0.64
Organizational	Organizational	0.83	0.83	0.83
	FinCausal	0.59	0.68	0.55
	SCITE	0.33	0.42	0.26

The objective is to understand the relation between the K–L divergence and the F1-score. With respect to the transfer learning task, from Table 2, we see that the performance of the model is much better when the model is fine-tuned on the FinCausal dataset. We got a macro average F1 score of 0.59. There is almost a 29% increase in the F1 score when Organizational data is used as a test rather than using the SCITE. Similarly, we can see a 13% increase in the F1 score when the model is fine-tuned on the Organizational data and tested on FinCausal rather than on the SCITE. The percentage increase between FinCausal as train and Organizational as the test is higher than Organizational as train and FinCausal as a test. This may be because of the number of training samples in Fincausal, which is higher than the number of training examples in Organizational data.

It means that the model performs better when there is more similarity between the vocabulary used in the train and test dataset and the number of samples is higher — confirming the intuitions. From Section 3, we know that the SCITE data is created from the web text, and the FinCausal data is annotated on the financial documents. To understand *how* the F1-score varies depending on the K–L divergence, Wasserstein distance, and Kolmogorov–Smirnov test p-values, we have plotted the dependencies in Fig. 3, in their simplest forms, as linear regression lines.

Linear regression R2 values help quantify these impressions. Thus K–L divergence predicts the performance of transfer learning with high accuracy as measured by R2 = [0.07462152], and confidence interval [−0.81916225 0.87093901]. The other two measures are not particularly good: surprisingly, Wasserstein distance gives R2 = [0.52912651], and the confidence interval [−0.03987017 0.78268854], the MMD gives R2 = [0.17625334], with the confidence interval [−0.92479103, 0.79202877], and the K-S test R2 = [0.40025979], with the confidence interval [0.48917194 0.38749281]. With a small number of points, we obtain wide confidence intervals. So, the results, even though confirming the observations of Hematialam and Zadrozny (2021) on medical transfer learning, have to be taken with a grain of salt. Nevertheless, they do suggest the higher predictive value of K–L divergence for this and perhaps similar tasks.

We can see that lower K–L divergence values predict higher F1 scores. The same is true for Wasserstein distance. The dependence between the computed p-values of Kolmogorov–Smirnov test and the F1 score is plotted in the bottom panel of Fig. 3, but the diagram does not seem informative — the high F1 values correspond to identical distributions.

We are also confirming this result using Spearman correlation. We compute the correlation between the predictive measures (K–L divergence, Wasserstein distance, and Kolmogorov–Smirnov test) and the F1 scores. (We used the Python Scipy library https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html). For example, the input to the Spearman correlation will be K–L-divergence (column 1 in Table 1) and F1 score (column 5 in Table 2).

```
KLd = [0, 0.942,0.906,0,0.771,0.286,0,0.279,0.336]
F1 = [0.73, 0.12,0.16, 0.92, 0.31,0.59, 0.78,0.39,0.26]
res = stats.spearmanr(KLd, F1)
print(res.statistic)
```

Thus we obtained the Spearman correlation of −0.94 between K–L divergence and the F1 score. This indicates a strong negative correlation between K–L divergence and F1 score, i.e., when the K–L divergence increases, the F1 score decreases. Similarly, we got the Spearman correlation of −0.69 between the Wasserstein distance and F1 score, indicating a negative correlation. With Kolmogorov–Smirnov test, we got the Spearman correlation of 0.69. This is because the higher the p-value, the higher the similarity between datasets which is opposite of the other two measures.

5.2. Confirming the results using SpanBERT and BERT

We performed additional transfer experiments using SpanBERT (Joshi et al., 2020), and BERT (Devlin et al., 2019). BERT and SpanBERT results also indicate the same interpretation as DistilBERT, i.e., we see a 32% increase in the F1 score when the model is fine-tuned on FinCausal data and Organizational data has used a test rather than SCITE. Similarly, with BERT and SpanBERT, there is a 12% and 29% increase in the F1 score when trained on Organizational data and tested on FinCausal rather than on SCITE.

With BERT, we got the Spearman correlation of −0.77 between K–L divergence and F1 score, −0.49 between Wasserstein distance and F1 score, and 0.50 between the KS test and F1 score. Similarly, with SpanBERT, we got the Spearman correlation of −0.94 between K–L divergence and F1 score, −0.73 between Wasserstein distance and F1 score, and 0.72 between the KS test and F1 score. The results of BERT and SpanBERT also indicate a strong negative correlation between predictive measures and F1 score. This suggests the strategy for data augmentation, where we choose examples that contribute to the decrease in the K–L divergence between the train and test.

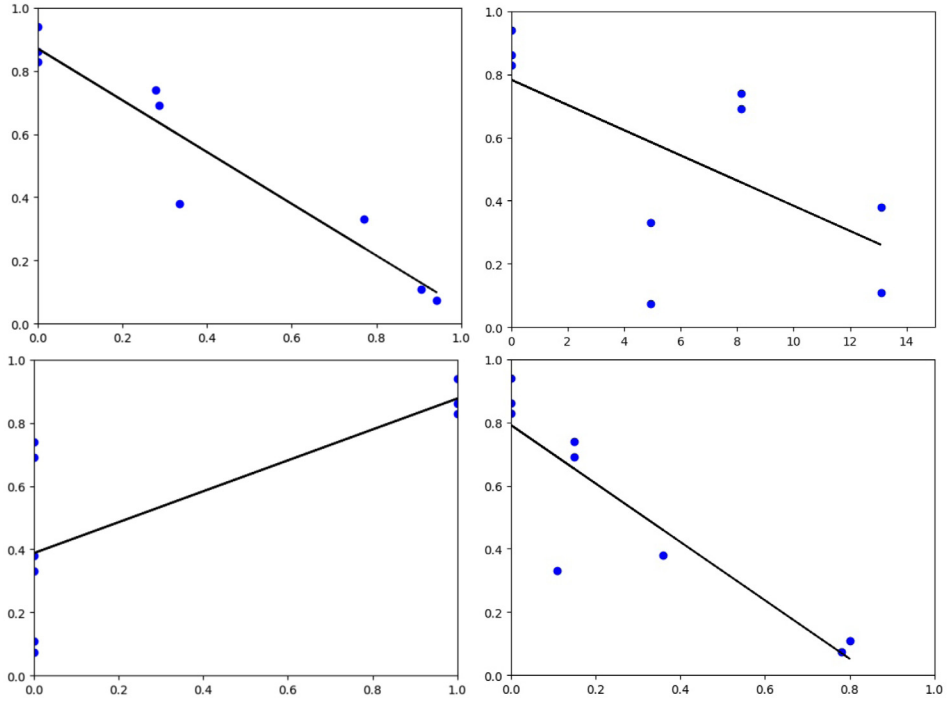


Fig. 3. Top Left Panel: A linear regression model approximating between the data points of the experiments; the K–L divergence between datasets is represented along the X-axis, and the F1-scores of the corresponding machine learning transfer experiments are shown represented along the Y-axis. **Top Right Panel:** The results of the same experiments using the Wasserstein distance instead of the K–L divergence. **Bottom Panel:** The results of the same experiments using the Kolmogorov–Smirnov test. **Bottom Left Panel:** The results of the same experiments using the Kolmogorov–Smirnov test. **Bottom Right Panel:** The results of the same experiments using the Maximum Mean Discrepancy (MMD).

5.3. Confirming the predictive measures with a few-shot transfer learning

We want to see if the strategy works. The results of comparing the predictive measures and the F1-score show that if there is some similarity between the training and the test data, there will be an increase in the performance, but it is not necessary that they should be in the same domain. In order to confirm these results, we added a small subset of the test data to the training data. Since all three models expressed a strong negative correlation and not much difference in performance, we report the few-shot transfer learning results by fine-tuning DistilBERT. We did four-fold cross-validation on the test data, and for each fold, we appended the smaller split to the training data as given in the equation below.

The results of this experiment when using DistilBERT are summarized in Table 5. There is a performance increase in all sets of the experiments. Specifically, when the model is fine-tuned on SCITE data and tested on FinCausal and Organizational data, there is approximately 60% to 63% increase in the performance. For the other set of experiments, when FinCausal and Organizational are used for training and tested on the other two datasets, the performance increase is relatively small. This may be because SCITE is an annotated corpus of text from the web, whereas FinCausal and Organizational are domain-specific datasets annotated on financial text. This method facilitates a method to choose a subset of data from the test data to increase the performance based on the K–L divergence value. For example, with the proposed method, when SCITE is used as a train, and FinCausal is used as a test, a small portion of the split, say 66 examples out of a total 265 examples in the test, are added to the SCITE data. Adding this reduces the K–L divergence value from 0.94 to 0.41, which gives approximately a 63% of increase in the F1 score (see Table 6).

Table 5
Summary of the few-shot transfer learning experiments. This table shows the performance of DistilBERT for causality extraction. Four-fold cross-validation was done. The smaller portion of the split is appended with the train data.

Train data	Test data	P	R	F1
SCITE	FinCausal	0.73	0.80	0.75
	Organizational	0.79	0.76	0.76
FinCausal	SCITE	0.42	0.48	0.43
	Organizational	0.79	0.77	0.77
Organizational	FinCausal	0.79	0.83	0.80
	SCITE	0.47	0.59	0.45

Table 6
Percentage increase between the results of DistilBERT when different domain data are used and when a small portion of the test data is added to the training data. There is an increase in the performance with all the sets of experiments. Specifically there is a huge increase in the performance when domain-specific data (FinCausal and Organizational) added to the general English data from web (SCITE).

Train data	Test data	% increase
SCITE	FinCausal	63%
	Organizational	60%
FinCausal	SCITE	12%
	Organizational	18%
Organizational	FinCausal	41%
	SCITE	19%

6. Discussion and future work

This work raises several questions. Regarding predictive value of K–L divergence and other tests, we only looked at word frequencies. For larger datasets, bigrams and trigrams could be added to the list of terms, and the tests could be repeated. In our domain such corpora do not exist, and bigrams and trigrams repeat infrequently. But it is beyond the scope of this article to investigate this larger version of the transfer problem.

Given the paucity of annotated data and the possible predictive value of K–L divergence, perhaps it can be used to guide a data augmentation strategy, that is choosing the texts that might have highest impact in reducing the differences between the distributions. It was

shown in Schlaefer et al. (2011) and Chu-Carroll et al. (2012) that a good strategy for data augmentation can have substantial impact on understanding meaning of texts in the context of question answering.

As to the choice of models, DistilBERT, it is perhaps still best performing model from the BERT family for the causality extraction task, based e.g. on our experiments with SpanBERT. The next natural question is: what about GPT? The use of new large language models like GPT-3.5, GPT-4, Llama and others remains an open problem. The issues in applying these models are discussed in a recent IEEE Spectrum article (https://spectrum.ieee.org/open-source-llm-not-open). We reported our experience with GPT-3 on Organizational data (Gopalakrishnan et al., 2023), where its performance was subpar compared to DistilBERT. And more recently, our preliminary experiments with GPT-4 with a few (1–16) shot learning, on a similar causality extraction task, we did not do better than using DistilBERT (report to appear later in August). Other researchers (Gao et al., 2023) confirm these observations. We hypothesize that these negative results are due to paucity of data annotated with causality markers — as we know these models required huge amount of data for training, and causality is not a simple linguistic concept, like e.g. dependency, and perhaps cannot be learned from raw text. Thus how to use LLMs for causality extraction seems to be an open problem, which we plan to continue to address.

7. Conclusion

In this article, we discussed the transfer learning (also called domain adaptation) performance of DistilBERT, a variant of BERT, for the causality extraction task. We reported results on pairwise transfer between three different datasets. We showed that the higher performance is correlated with lower difference in distributions of word frequencies in the datasets. And we quantified these differences using three measures: K–L divergence, Wasserstein distance, and Kolmogorov–Smirnov test. We estimated the predictive values of three tests for transfer learning (in this domain). We report K–L divergence performed the best.

For our experiments, we use three datasets Organizational (Gopalakrishnan et al., 2023), annotated on financial text; SCITE (Li et al., 2021b), annotated on texts from the web; and FinCausal (Mariko et al., 2022), annotated on the financial news articles. Even though Organizational data and FinCausal data are created from financial texts, both of these datasets have different distributions, which was shown to lead to different success rate in transfer learning. Since our work is based on information theory, we hypothesize that this transfer learning approach should work for other domains.

We also discussed preliminary experiments with new large language models in the GPT family, and plans to use these and other large language models. It is yet to be seen which of them can support transfer learning, or guarantee out-of-the-box high performance on causality extraction tasks.

CRediT authorship contribution statement

Seethalakshmi Gopalakrishnan: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Formal analysis. **Victor Zitian Chen:** Supervision, Project administration. **Wenwen Dou:** Supervision, Project administration. **Wlodek Zadrozny:** Writing – review & editing, Supervision, Project administration, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We appreciate the referees feedback, which helped us improve the paper.

Funding

This research was partly funded by the National Science Foundation (NSF), USA, grant number 2141124.

References

Al Kuwatly, H., Wich, M., Groh, G., 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In: Proceedings of the Fourth Workshop on Online Abuse and Harms. pp. 184–190.

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al., 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Bui, Q.-C., Nualláin, B.Ó., Boucher, C.A., Sloot, P.M., 2010. Extracting causal relations on HIV drug resistance from literature. BMC Bioinform. 11, 1–11.

Chen, X., Li, Q., Li, Z., Xie, H., Wang, F.L., Wang, J., 2022. A reinforcement learning based two-stage model for emotion cause pair extraction. IEEE Trans. Affect. Comput..

Chen, X., Li, Z., Wang, Y., Xie, H., Wang, J., Li, Q., 2023. Recognizing conditional causal relationships about emotions and their corresponding conditions. arXiv preprint arXiv:2311.16579.

Chen, X., Sun, Y., Athiwaratkun, B., Cardie, C., Weinberger, K., 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. Trans. Assoc. Comput. Linguist. 6, 557–570.

Chu-Carroll, J., Fan, J., Schlaefer, N., Zadrozny, W., 2012. Textual resource acquisition and engineering. IBM J. Res. Dev. 56 (3.4), 4.1–4.11.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A., 2017. Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364.

Dang, H.A., 2021. A study on extracting Cause-Effect relations and these application for Why-question answering.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. http://dx.doi.org/10.18653/v1/N19-1423, URL: https://www.aclweb.org/anthology/N19-1423.

Ding, Z., He, H., Zhang, M., Xia, R., 2019b. From independent prediction to reordered prediction: Integrating relative position and global label information to emotion cause identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 6343–6350.

Ding, Y., Tang, J., Guo, F., 2019a. Identification of drug-side effect association via multiple information integration with centered kernel alignment. Neurocomputing 325, 211–224.

Gao, J., Ding, X., Qin, B., Liu, T., 2023. Is ChatGPT a good causal reasoner? A comprehensive evaluation. arXiv preprint arXiv:2305.07375.

Garcia, D., EDF-DER, IMA-TIEM, 1997. COATIS, an NLP system to locate expressions of actions connected by causality links. In: Knowledge Acquisition, Modeling and Management: 10th European Workshop, EKAW'97 Sant Feliu de Guixols, Catalonia, Spain October 15–18, 1997 Proceedings 10. Springer, pp. 347–352.

Gibbs, A.L., Su, F.E., 2002. On choosing and bounding probability metrics. Int. Stat. Rev. 70 (3), 419–435.

Girju, R., 2003. Automatic detection of causal relations for question answering. In: Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering. pp. 76–83.

Gopalakrishnan, S., Chen, V.Z., Dou, W., Hahn-Powell, G., Nedunuri, S., Zadrozny, W., 2023. Text to causal knowledge graph: A framework to synthesize knowledge from unstructured business texts into causal graphs. Information 14 (7), 367.

Gu, J., Qian, L., Zhou, G., 2016. Chemical-induced disease relation extraction with various linguistic features. Database 2016.

Hassanzadeh, O., Bhattacharjya, D., Feblowitz, M., Srinivas, K., Perrone, M., Sohrabi, S., Katz, M., 2019. Answering binary causal questions through large-scale text mining: An evaluation using cause-effect pairs from human experts. In: IJCAI. pp. 5003–5009.

Hematialam, H., Zadrozny, W.W., 2021. Identifying condition-action statements in medical guidelines: Three studies using machine learning and domain adaptation.

Hendrickx, I., Kim, S.N., Kozareva, Z., Nakov, P., Séaghdha, D.O., Padó, S., Pennacchiotti, M., Romano, L., Szpakowicz, S., 2019. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. arXiv preprint arXiv:1911.10422.

Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., Levy, O., 2020. Spanbert: Improving pre-training by representing and predicting spans. Trans. Assoc. Comput. Linguist. 8, 64–77.

Kang, N., Singh, B., Bui, C., Afzal, Z., van Mulligen, E.M., Kors, J.A., 2014. Knowledge-based extraction of adverse drug events from biomedical text. BMC Bioinformatics 15 (1), 1–8.

Khetan, V., Ramnani, R., Anand, M., Sengupta, S., Fano, A.E., 2020. Causal BERT: Language models for causality detection between events expressed in text. arXiv preprint arXiv:2012.05453.

Khoo, C.S., Myaeng, S.H., Oddy, R.N., 2001. Using cause-effect relations in text to improve information retrieval precision. Inf. Process. Manage. 37 (1), 119–145.

Kyriakakis, M., Androutsopoulos, I., Saudabayev, A., et al., 2019. Transfer learning for causal sentence detection. arXiv preprint arXiv:1906.07544.

Li, X., 2022. Causal domain adaptation for information extraction from complex conversations. In: European Semantic Web Conference. Springer, pp. 189–198.

Li, D., Fei, H., Ren, S., Li, P., 2021a. A deep decomposable model for disentangling syntax and semantics in sentence representation. In: Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 4300–4310.

Li, Z., Li, Q., Zou, X., Ren, J., 2021b. Causality extraction based on self-attentive BiLSTM-CRF with transferred embeddings. Neurocomputing 423, 207–219.

Li, F., Zhang, M., Fu, G., Ji, D., 2017. A neural joint model for entity and relation extraction from biomedical text. BMC Bioinform. 18 (1), 1–11.

Lyu, C., Ji, T., Sun, Q., Zhou, L., 2022. DCU-Lorcan at FinCausal 2022: Span-based causality extraction from financial documents using pre-trained language models. In: Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022. pp. 116–120.

Manning, C.D., 2009. An Introduction to Information Retrieval. Cambridge University Press.

Mariko, D., Abi Akl, H., Trottier, K., El-Haj, M., 2022. The financial causality extraction shared task (FinCausal 2022). In: Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022. pp. 105–107.

Martin, J.H., 2009. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Pearson/Prentice Hall.

Miller, T., Laparra, E., Bethard, S., 2021. Domain adaptation in practice: Lessons from a real-world information extraction pipeline. In: Proceedings of the Second Workshop on Domain Adaptation for NLP. pp. 105–110.

Mozafari, M., Farahbakhsh, R., Crespi, N., 2020. A BERT-based transfer learning approach for hate speech detection in online social media. In: Complex Networks and their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and their Applications COMPLEX NETWORKS 2019 8. Springer, pp. 928–940.

Pakray, P., Gelbukh, A., 2014. An open-domain cause-effect relation detection from paired nominals. In: Nature-Inspired Computation and Machine Learning: 13th Mexican International Conference on Artificial Intelligence, MICAI 2014, Tuxtla Gutiérrez, Mexico, November 16-22, 2014. Proceedings, Part II 13. Springer, pp. 263–271.

Pan, S.J., Yang, Q., 2009. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22 (10), 1345–1359.

Pan, S.J., Yang, Q., 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22 (10), 1345–1359.

Peng, B., Chersoni, E., Hsu, Y.-Y., Huang, C.-R., 2021. Is domain adaptation worth your investment? Comparing BERT and FinBERT on financial tasks. In: Proceedings of the Third Workshop on Economics and Natural Language Processing. pp. 37–44.

Peng, Y., Yan, S., Lu, Z., 2019. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. arXiv preprint arXiv:1906.05474.

Radinsky, K., Davidovich, S., Markovitch, S., 2012. Learning causality for news events prediction. In: Proceedings of the 21st International Conference on World Wide Web. pp. 909–918.

Rietzler, A., Stabinger, S., Opitz, P., Engl, S., 2019. Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification. arXiv preprint arXiv:1908.11860.

Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Schlaefer, N., Chu-Carroll, J., Nyberg, E., Fan, J., Zadrozny, W., Ferrucci, D., 2011. Statistical source expansion for question answering. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management pp. 345–354.

Sobrino, A., Puente, C., Olivas, J.A., 2014. Extracting answers from causal mechanisms in a medical document. Neurocomputing 135, 53–60.

Sun, S., Shi, H., Wu, Y., 2015. A survey of multi-source domain adaptation. Inf. Fusion 24, 84–92.

Sun, C., Yang, Z., 2019. Transfer learning in biomedical named entity recognition: an evaluation of BERT in the PharmaCoNER task. In: Proceedings of the 5th Workshop on BioNLP Open Shared Tasks. pp. 100–104.

Weiss, K., Khoshgoftaar, T.M., Wang, D., 2016. A survey of transfer learning. J. Big Data 3 (1), 1–40.

Xia, R., Ding, Z., 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. arXiv preprint arXiv:1906.01267.

Yang, J., Han, S.C., Poon, J., 2022. A survey on extraction of causal relations from natural language text. Knowl. Inf. Syst. 64 (5), 1161–1186.

Zhang, Y., Qi, P., Manning, C.D., 2018. Graph convolution over pruned dependency trees improves relation extraction. arXiv preprint arXiv:1809.10185.

Zoph, B., Ghiasi, G., Lin, T.-Y., Cui, Y., Liu, H., Cubuk, E.D., Le, Q., 2020. Rethinking pre-training and self-training. Adv. Neural Inf. Process. Syst. 33, 3833–3845.