Causalité 2 – Création d'embeddings de textes pour l'inférence causale

Séminaire TALia du 25/11/2022

Séminaires TALia pertinents
- 21 – Transformers (1)
- 56 – Introduction aux VAE
- 65 – Causalité 1 - L'approche SCM de Pearl

PMLR

Adapting Text Embeddings for Causal Inference

Victor Veitch, Dhanya Sridhar, David Blei *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, PMLR 124:919-928, 2020.

vidéo

arXiv > cs > arXiv:2109.00725

Computer Science > Computation and Language

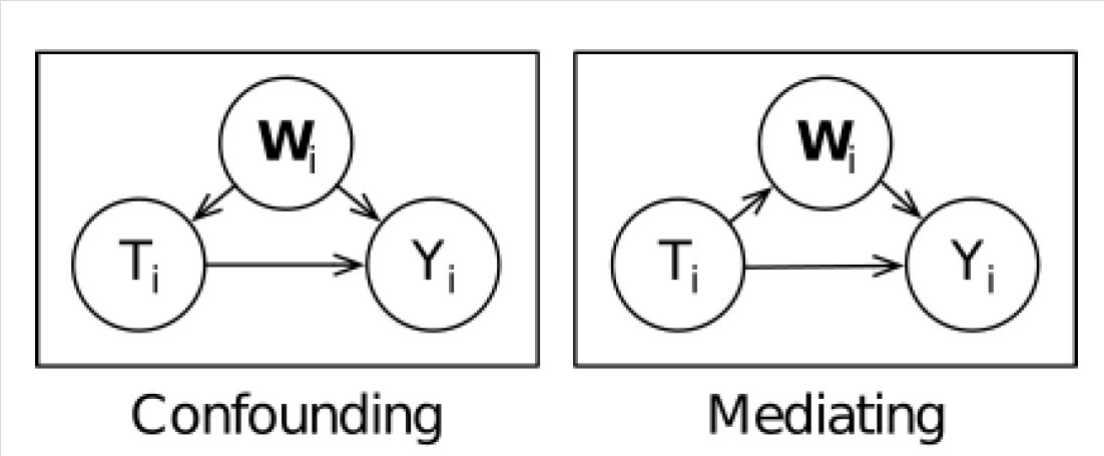
[Submitted on 2 Sep 2021 (v1), last revised 30 Jul 2022 (this version, v2)]

Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, Diyi Yang

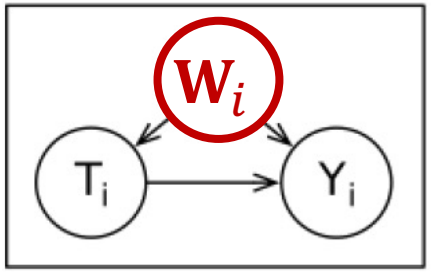
Plan

1. Deux exemples d'inférence causale avec du texte
2. Une réduction dimensionnelle qui permet l'identification
3. Deux implémentations : C-BERT et C-ATM (LDA)
4. Un jeu de données semi-synthétique
5. Une application sur des données réelles
6. Deux questions ouvertes



Deux exemples simples d'inférence causale à partir de textes

Exemple n°1
On souhaite examiner l'impact causal de l'inclusion d'un théorème dans un article scientifique sur ses chances d'être publié. L'acceptation dépend de critères liés au caractère plus ou moins technique du sujet. Ce sujet peut être inféré de l'abstract.

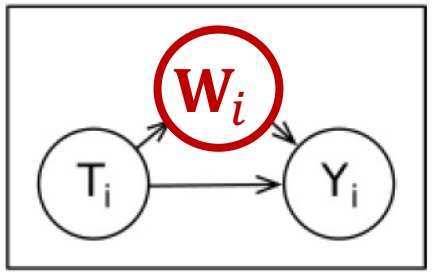


Confounding

back-door rule

- $T \in \{0,1\}$ présence d'un théorème dans l'article
- **W** le texte de l'abstract
- $Y \in \{0,1\}$ acceptation de l'article

Exemple n°2
On souhaite déterminer l'impact causal de l'affichage d'une icône de genre sur la popularité d'une post dans un réseau social.



Mediating

cause sans parents

- $T \in \{H, F\}$ affichage du genre
- **W** le texte du post
- $Y \in \mathbb{R}$ la popularité du post

Point commun : Le résultat Y dépend simultanément d'un « confounder » textuel **W** et d'un « traitement » T.

Deux mesures de l'effet causal

Observations : $O_i := (Y_i, T_i, \mathbf{W}_i) \sim P, \quad i = 1, ..., n$

$Z_i := f(\mathbf{W}_i)$ représentation/réduction dimensionnelle du texte.

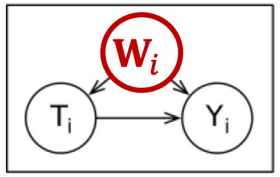
S. 65 – Causalité 1

Théorème (*back-door adjustment*) : Si un ensemble de variables **Z** satisfait la back-door rule par rapport a X et Y alors l'effet causal $P(Y|\text{do}(X))$ est **identifiable** au moyen de :

$$p(y|\text{do}(X = x)) = \sum_{\mathbf{z}} p(y|x, \mathbf{z})\, p(\mathbf{z})$$

Average Treatment Effect on the Treated

$$\text{ATT} := \mathbb{E}_Y[Y|\text{do}(T=1), T=1] - \mathbb{E}_Y[Y|\text{do}(T=0), T=1]$$



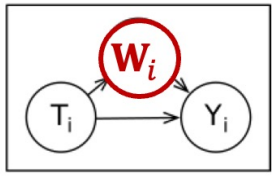
Confounding

back-door rule ✓

$$\overset{\text{back door rule}}{=} \mathbb{E}_Z \left[\underbrace{\mathbb{E}_Y[Y|Z, T=1]}_{:= Q(Z, T=1)} - \underbrace{\mathbb{E}_Y[Y|Z, T=0]}_{:= Q(Z, T=0)} \,\middle|\, T = 1 \right]$$

Natural Direct Effect

$$\text{NDE} := \mathbb{E}_Z \left[\mathbb{E}_Y[Y|Z, \text{do}(T=1)] - \mathbb{E}_Y[Y|Z, \text{do}(T=0)] \,\middle|\, T = 1 \right]$$



Mediating

cause sans parents ✓

$$\overset{\text{PA}(T)=\varnothing}{=} \mathbb{E}_Z \left[\mathbb{E}_Y[Y|Z, (T=1)] - \mathbb{E}_Y[Y|Z, (T=0)] \,\middle|\, T = 1 \right]$$

DE : Espérance de la variation de l'effet Y lorsqu'on maintient fixe les variables médiatrices Z affectées par le traitement T.

NDE : Pondération de **DE** avec une distribution de Z qui est celle que l'on observe lorsque le traitement est administré.

Deux estimateurs de l'effet causal

Estimateurs de base

$$
\begin{aligned}
\mathbb{E}_Y[Y|Z=z,T=t] := Q(z,t) &\simeq \hat{Q}(z,t) \\
p(T=1|z) := g(z) &\simeq \hat{g}(z) \\
p(T=1) &\simeq \frac{1}{n}\sum_{i=1}^{n} t_i \\
\mathbb{E}_Z[f(Z)] &\simeq \frac{1}{n}\sum_{i=1}^{n} \hat{f}(z_i) \\
\mathbb{E}_Z[f(Z)|T=1] &\simeq \sum_{i=1}^{n} t_i\, \hat{f}(z_i) \Big/ \sum_{i=1}^{n} t_i
\end{aligned}
$$

Estimateur Q-only

$$
\begin{aligned}
\text{ATT} &= \mathbb{E}_Z[Q(Z,T=1) - Q(Z,T=0)|T=1] \\
&\simeq \frac{\sum_{i=1}^{n} t_i \left[\hat{Q}(z_i,T=1) - \hat{Q}(z_i,T=0)\right]}{\sum_{i=1}^{n} t_i} =: \widehat{\text{ATT}}^{Q}
\end{aligned}
$$

Estimateur plugin

$$
\begin{aligned}
\text{ATT} &= \mathbb{E}_Z[Q(Z,T=1) - Q(Z,T=0)|T=1] \\
&= \int [Q(z,T=1) - Q(z,T=0)]\, p(z|T=1)\mathrm{d}z \\
&= \int [Q(z,T=1) - Q(z,T=0)]\, \frac{\overbrace{p(T=1|z)}^{:=g(z)}\, p(z)}{p(T=1)}\mathrm{d}z \\
&= \frac{\mathbb{E}_Z[[Q(Z,T=1) - Q(Z,T=0)]\, g(Z)]}{p(T=1)} \\
&\simeq \frac{\sum_{i=1}^{n} [\hat{Q}(z_i,T=1) - \hat{Q}(z_i,T=0)]\, \hat{g}(z_i)}{\sum_{i=1}^{n} t_i} =: \widehat{\text{ATT}}^{\text{plugin}}
\end{aligned}
$$

Une réduction dimensionnelle qui préserve l'identifiabilité

But : trouver une réduction dimensionnelle $z = f(\mathbf{w})$ des suites de mots $\mathbf{w}$ **suffisante** (= qui conserve assez d'information) pour que l'effet causal soit **identifiable**.

$$\begin{aligned}\text{ATT} &= \mathbb{E}_{\mathbf{W}}\left[Q(\mathbf{W}, T=1) - Q(\mathbf{W}, T=0) \,\middle|\, T=1\right] \\ &= \mathbb{E}_{Z}\left[\bar{Q}(\underbrace{\lambda(\mathbf{W})}_{:=Z}, T=1) - \bar{Q}(\underbrace{\lambda(\mathbf{W})}_{:=Z}, T=0) \,\middle|\, T=1\right]\end{aligned}$$

Intuition n°1 :

Comme l'effet y correspond à une **évaluation humaine** on choisit une représentation z qui encode le sens de la phrase $\mathbf{w}$.

On peut essayer p.ex. :

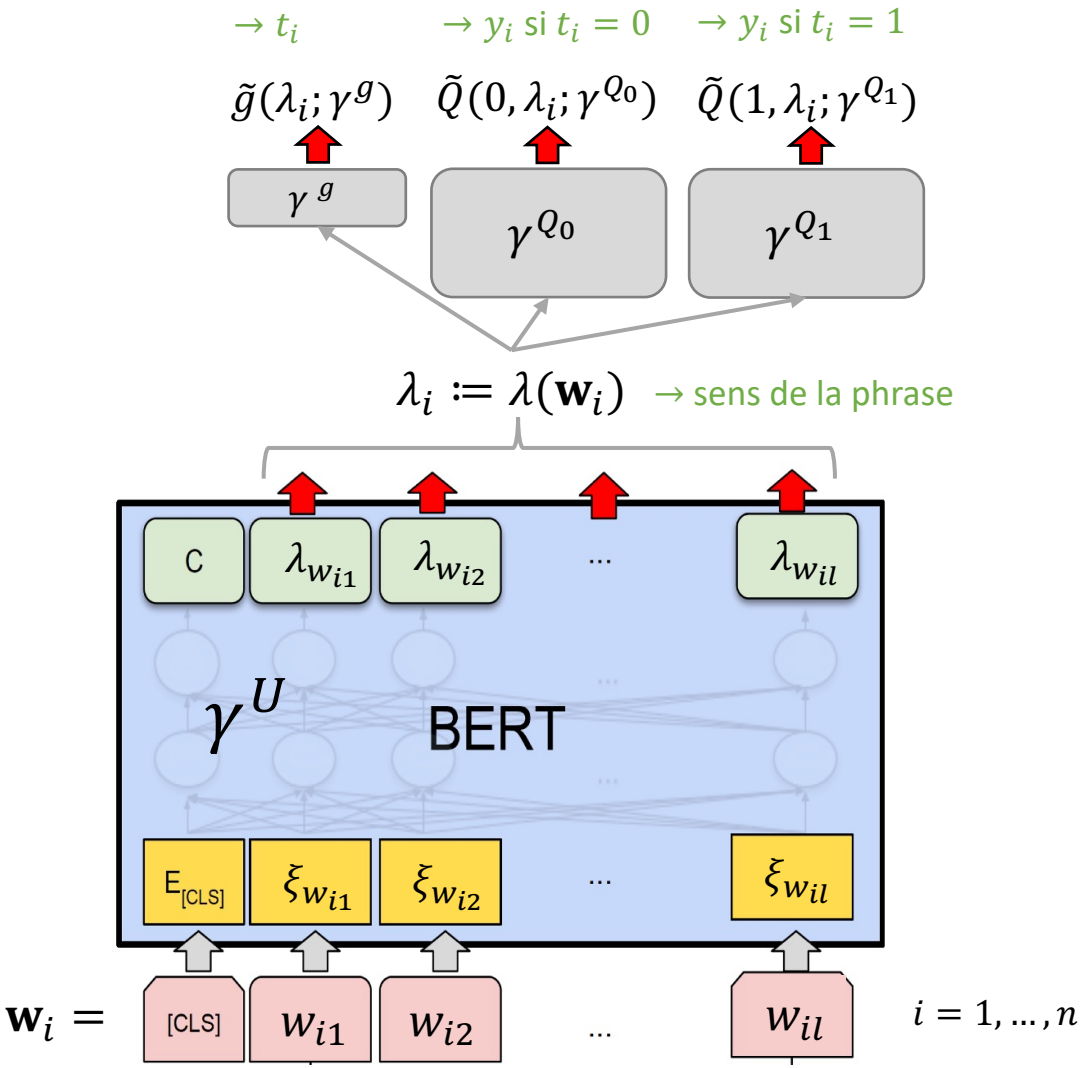
- les **embeddings** $z(\mathbf{w}) = \lambda(\mathbf{w})$ en sortie d'un BERT
- les **thèmes** $z(\mathbf{w}) = \theta(\mathbf{w})$ découverts par LDA ou ATM

Intuition n°2 :

L'effet causal sera identifiable à condition que z contienne assez d'information pour prédire à la fois l'**espérance de l'effet** $\hat{Q}(z,t)$ et le **propensity score** $\hat{g}(z)$.

$$g(\mathbf{w})\, p(\mathbf{w})\, \mathrm{d}\mathbf{w} = \bar{g}(z(\mathbf{w}))\, \bar{p}(z(\mathbf{w}))\, \mathrm{d}z$$
$$Q(t,\mathbf{w})\, p(\mathbf{w})\, \mathrm{d}\mathbf{w} = \bar{Q}(t, z(\mathbf{w}))\, \bar{p}(z(\mathbf{w}))\, \mathrm{d}z$$

Réduction dimensionnelle n°1 : C–BERT



Entraînement : fine-tuning de BERT

$$\hat{\xi}, \hat{\gamma} := \arg\min_{\xi,\gamma} \frac{1}{n} \sum_{i=1}^{n} L(\mathbf{w}_i, t_i, y_i ; \xi, \gamma)$$

$$L(\mathbf{w}_i, t_i, y_i ; \xi, \gamma) = \underbrace{L_{\text{MLM}}(\mathbf{w}_i ; \xi, \gamma^U)}_{\text{préentraînement}} + \underbrace{L_g(\mathbf{w}_i, t_i ; \xi, \gamma^g) + L_Q(\mathbf{w}_i, t_i, y_i ; \xi, \gamma^{Q_{t_i}})}_{\text{fine tuning}}$$

$$L_g(\mathbf{w}_i, t_i ; \xi, \gamma^g) := \text{CrossEnt}\left(t_i, \tilde{g}(\lambda_i ; \gamma^g)\right)$$

$$L_Q(\mathbf{w}_i, t_i, y_i ; \xi, \gamma^{Q_{t_i}}) := \left(y_i - \tilde{Q}(t_i, \lambda_i, \gamma^{Q_{t_i}})\right)^2$$

Estimation de l'effet : insérer les prédictions $\tilde{g}(\lambda_i)$ et $\tilde{Q}(t_i, \lambda_i)$ sur un ensemble de validation dans $\widehat{\text{ATT}}$.

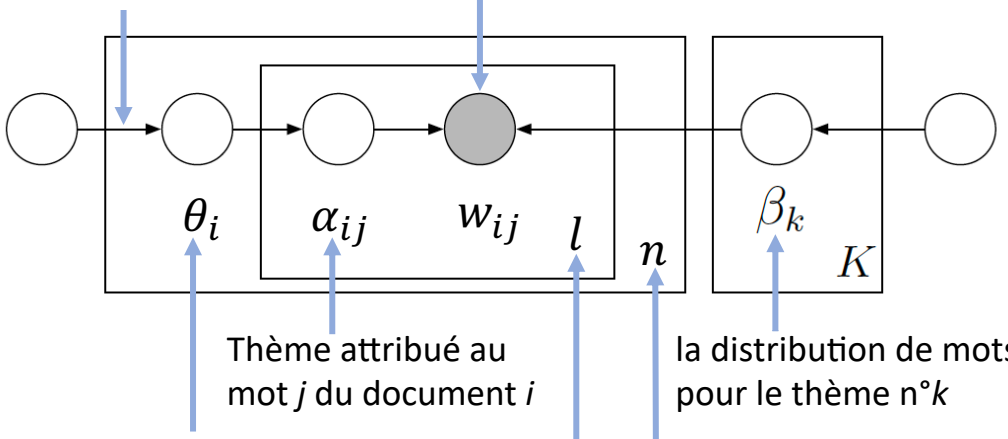
Réduction dimensionnelle n°2 : C–ATM

ATM : modèle génératif de type LDA

Prior sur proportions de thèmes

$\theta_i \sim p(\theta_i)$
$:= \text{softmax}(\mathcal{N}(0, \mathbb{I}_k))$

*j*ème mot observé dans le document *i*



Thème attribué au mot *j* du document *i*

Proportion des thèmes pour le document *i*.
Ce sont les variables latentes.

la distribution de mots pour le thème n°*k*

nb de documents

nb de mots par document

Variables observées : $\mathbf{w}_i$

S. 56 – Introduction aux VAE

Maximum likelihood sur les $\mathbf{w}_i$ par inférence variationnelle

$$-L_{\text{ELBO}}(\mathbf{w}_i; \beta, \eta) := \mathbb{E}_{q(\theta|\mathbf{w}_i;\eta)}[\, p(\mathbf{w}_i|\theta; \beta)] - \text{KL}[q(\theta|\mathbf{w}_i; \eta)\|p(\theta)]$$

Embedding dans l'espace latent :

$$\mathbf{w}_i \to q(\theta|\mathbf{w}_i; \eta) = \mathcal{N}(\mu(\mathbf{w}_i), \sigma(\mathbf{w}_i))$$

C'est l'analogue de : $\mathbf{w}_i \to \lambda_i = \lambda(\mathbf{w}_i)$

Fonction de coût complète

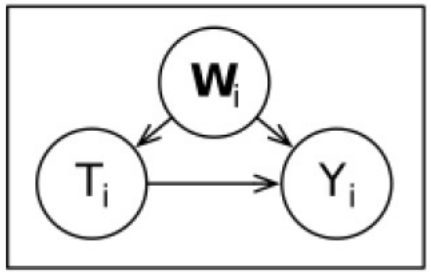
$$\begin{aligned} L(\mathbf{w}_i; \eta, \beta, \gamma) := \quad & L_{\text{ELBO}}(\mathbf{w}_i; \beta, \eta) \\ + \quad & \mathbb{E}_{q(\theta|\mathbf{w}_i;\eta)}[\text{CrossEnt}\,(t_i, \tilde{g}(\theta\,; \gamma^g))] \\ + \quad & \mathbb{E}_{q(\theta|\mathbf{w}_i;\eta)}\left[\left(y_i - \tilde{Q}(t_i, \theta, \gamma^{Q_{t_i}})\right)^2\right] \end{aligned}$$

Un train set semi-synthétique pour valider l'approche

Difficulté fondamentale : Valider la méthode d'estimation causale est délicat car il n'existe pas de train set pour lequel on connaisse l'effet causal sur un texte.

Solution : Créer un train set **semi-synthétique**, qui reflète le monde réel en définissant une **réponse y_i artificielle** qui dépend à la fois du traitement t_i et d'un *confounder* $z(\mathbf{w}_i)$ bien défini et corrélé au texte $\mathbf{w}_i$.

Décision d'acceptation **fictive** : $$Y_i \sim \text{Bernoulli}(\sigma(0.25t_i + b_1(\pi(\tilde{z}_i) - 0.2)))$$



Confounding

b_1 contrôle l'importance du *confounder*

$\pi(\tilde{z}_i)$ est le vrai *propensity score*

Données réelles

t_i = présence d'un théorème dans le corps de l'article.

$\tilde{z}_i$ = le titre contient un buzzword comme '*deep*', '*neural*', '*adversarial*',…

Validation de la démarche sur le train set synthétique

Omission de la modélisation de la langue
via la tâche MLM dans BERT p.ex.

$$\cancel{L_{\text{MLM}}(\mathbf{w}_i\,;\xi,\gamma^U)}$$
$$+L_g(\mathbf{w}_i,t_i\,;\xi,\gamma^g)+L_Q(\mathbf{w}_i,t_i,y_i\,;\xi,\gamma^{Q_{t_i}})$$

Omission de la supervision de l'apprentissage
des embeddings $\lambda_i := \lambda(\mathbf{w}_i)$

$$L_{\text{MLM}}(\mathbf{w}_i\,;\xi,\gamma^U)$$
$$+\cancel{L_g(\mathbf{w}_i,t_i\,;\xi,\gamma^g)}+\cancel{L_Q(\mathbf{w}_i,t_i,y_i\,;\xi,\gamma^{Q_{t_i}})}$$

(a) Language Modeling Helps

Dataset:	Reddit (NDE)	PeerRead (ATT)
Ground truth	1.00	0.06
Unadjusted	1.24	0.14
NN $\hat{\psi}^Q$	1.17	0.10
NN $\hat{\psi}^{\text{plugin}}$	1.17	0.10
BERT (sup. only) $\hat{\psi}^Q$	0.93	0.19
BERT (sup. only) $\hat{\psi}^{\text{plugin}}$	1.17	0.18
C-ATM $\hat{\psi}^Q$	1.16	0.10
C-ATM $\hat{\psi}^{\text{plugin}}$	1.13	0.10
C-BERT $\hat{\psi}^Q$	1.07	0.07
C-BERT $\hat{\psi}^{\text{plugin}}$	1.15	0.09

(b) Supervision Helps

Dataset:	Reddit (NDE)	PeerRead (ATT)
Ground truth	1.00	0.06
Unadjusted	1.24	0.14
BOW $\hat{\psi}^Q$	1.17	0.13
BOW $\hat{\psi}^{\text{plugin}}$	1.18	0.14
BERT $\hat{\psi}^Q$	-15.0	-0.25
BERT $\hat{\psi}^{\text{plugin}}$	-14.1	-0.28
LDA $\hat{\psi}^Q$	1.20	0.07
LDA $\hat{\psi}^{\text{plugin}}$	1.20	0.09
ATM $\hat{\psi}^Q$	1.17	0.08
ATM $\hat{\psi}^{\text{plugin}}$	1.17	0.08

Application à des données réelles

L'application de cette méthode d'évaluation de l'effet causal de la présence de *buzzwords* ou d'un théorème montre que le **score non ajusté qui semble révéler un effet causal s'explique** en réalité **par l'influence des *confounders***.

	`buzzy`	`theorem`
Unadjusted	0.08 ± 0.01	0.21 ± 0.01
C-BERT $\hat{\psi}^{Q}$	-0.03 ± 0.01	0.16 ± 0.01
C-BERT $\hat{\psi}^{\text{plugin}}$	-0.02 ± 0.01	0.18 ± 0.02

Deux questions ouvertes

- Étendre l'approche à des **situations où le traitement** (dont on veut analyser l'impact causal) **est contenu dans le texte**. Dans cet article le traitement est externe, puisqu'il s'agit de la présence ou de l'absence d'un théorème dans le corps de l'article.

- Généraliser la méthode d'évaluation par train set semi-synthétique pour imaginer une **méthode de benchmarking** d'effets causal.