DEEL

DEpendable & Explainable Learning

IVADO

IRT SAINT EXUPÉRY

CRIAQ

ANITI Université Fédérale Toulouse Midi-Pyrénées

Institut intelligence et données

DEEL

DEpendable & Explainable Learning

L’explicabilité dans le NLP : identifier les biais dans les gros modèles de langue

Groupe Statistique Mathématique – 16/01/2025

fanny.jourdan@irt-saintexupery.com

IVADO

IRT SAINT EXUPÉRY

CRIAQ

ANITI Université Fédérale Toulouse Midi-Pyrénées

Institut intelligence et données

Plan

1. L’XAI pour la détection de biais
2. Méthodes d’attributions
   a. Basées sur les perturbations
   b. Basées sur les gradients
   c. Basées sur la structure interne
   d. Basées sur les valeurs de Shapley
3. Méthodes basées sur les concepts
   a. Avec des concepts supervisés
   b. Avec des concepts non-supervisés
4. Méthodes Contrastives et Contrefactuelles

DEEL

DEpendable & Explainable Learning

1. L'XAI dans la détection de biais pour les modèles NLP

IVADO

IRT SAINT EXUPÉRY

CRIAQ

ANITI Université Fédérale Toulouse Midi-Pyrénées

Institut intelligence et données

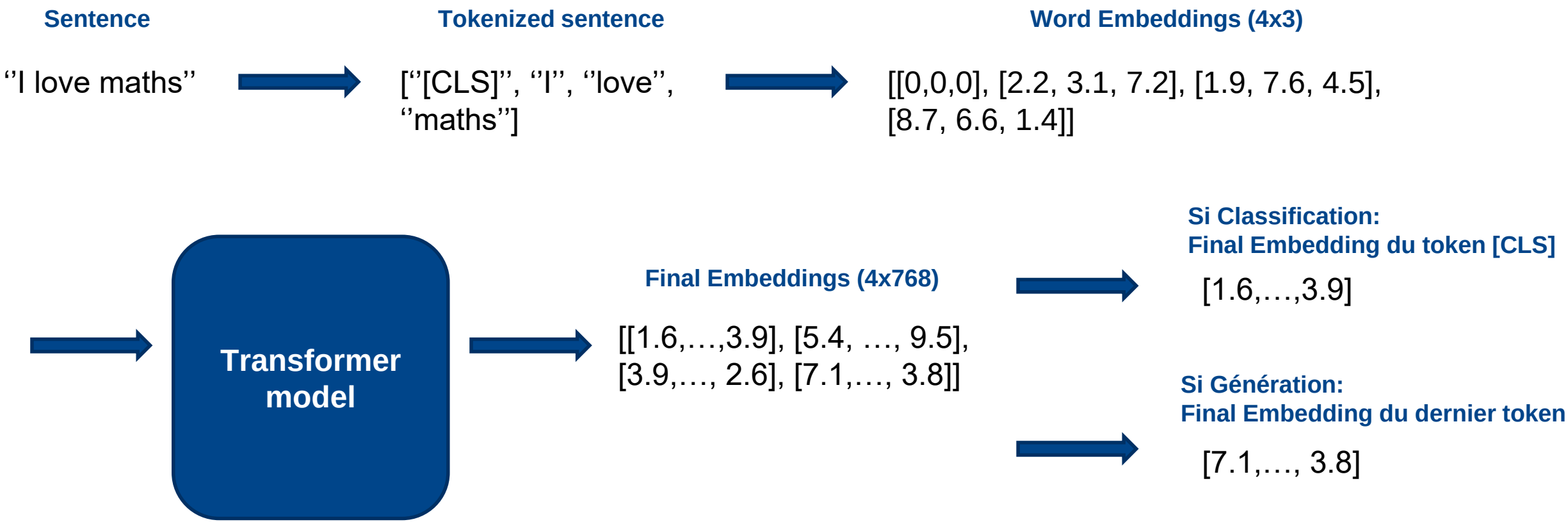
Qu’est ce que l’Explicabilité?

L’explicabilité vise à rendre les décisions des modèles compréhensibles pour les humains. Cela permet de répondre à des questions comme:

- *Pourquoi le modèle à fait cette prédiction?*
- *Quels caractéristiques sont les plus importants pour ce modèle?*
- *Quels sont les concepts appris par le modèle, et comment sont-ils utilisés dans les décisions ?*

Une méthode d’explicabilité efficace permet d’identifier, de comprendre et de corriger les biais dans les modèles, contribuant ainsi à des décisions plus équitables et responsables.

Enjeux spécifiques des modèles de NLP



Enjeux spécifiques des modèles de NLP

Classification

Final Embedding du token [CLS]

[1.6,…,3.9]

Logits sur les classes finales (5)

[0.2, 0.45, 0.001, 0.23, **0.89**]

Ici on prédira la classe numéro 5

Génération

Final Embedding du dernier token

[7.1,…, 3.8]

Logits tous les mots du vocabulaire (plus de 20k)

[0.34, ……., 0.67]

Ici on prédira le mot numéro 8561

Enjeux spécifiques des modèles de NLP

- Projette les word embeddings du début dans une espace optimale.
- Les projections finales de chaque mot contiennent des informations de tous les autres mots de la phrase.
- Non linéaire.
- Des milliards de paramètres.

Pour aller plus loin: https://jalammar.github.io/illustrated-transformer/

Overview des méthodes d'XAI en NLP

DEEL
DEpendable & Explainable Learning

Attribution methods

Perturbation-based attribution

→ Occlusion, DeconvNet *(Zeiler and Fergus, 2014)*
→ LIME *(Ribeiro et al., 2016)*
→ Sobol-based *(Fel et al., 2021)*
→ HSIC-based *(Novelo et al., 2022)*
→ ReAGent *(Zhao and Shan, 2024)*
...

Gradient-based attribution

→ Saliency maps, Input × Gradient *(Simonyan et al., 2014)*
→ Integrated Gradients *(Sundararajan et al., 2017)*
→ DeepLIFT *(Shrikumar et al., 2017)*
...

Internal-based attribution

→ Attention Weight *(Bahdanau et al., 2014)*
→ Attention Flow *(Abnar and Zuidema, 2020)*
...

SHAP techniques

→ SHAP, KernelSHAP, DeepSHAP *(Lundberg and Lee, 2017)*
→ C- and L-Shapley *(Chen et al., 2018)*
→ Integrated Hessians *(Janizek et al., 2021)*
...

Concept-based methods

Unsupervised concepts

→ ACE *(Ghorbani et al., 2019)*
→ CRAFT *(Fel et al., 2023)*
→ COCKATIEL *(Jourdan et al, 2023)*
...

Supervised concepts

→ TCAV *(Kim et al., 2018)*
→ *CBMs (Koh et al., 2020)*
→ *PCBMs (Yuksekgonul et al., 2022)*
...

SHAP x Unsupervised concepts

→ ConceptSHAP *(Yeh et al., 2020)*
...

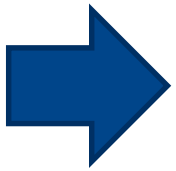
Rationalization

→ HardKuma *(Jasmijn et al., 2020)*
→ FRESH *(Jain et al., 2020)*
...

Supervised concepts x Rationalization

→ RNP *(Lei et al., 2016)*
→ RNP-3P *(Yu et al., 2019)*
→ InvRAT *(Chang et al., 2020)*
→ ConRAT *(Antognini and Faltings, 2021)*
...

Attribution pour une tache de classification

J’ai adoré ce film ! C’est le meilleur film que j’ai vu.  Avis Positif 

Heatmap de l’importance des mots de l’exemple pour la classe « positive »

Attribution pour une tache de génération

L’enseignante adore aider ses étudiants

The teacher loves

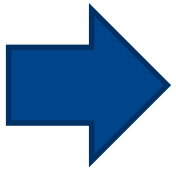
Heatmap de l’importance des mots précédents pour la génération du mot « loves »

Détection de biais avec l’attribution

Elle travaille à l’hôpital de Perpignan depuis 3 ans.

Les patients qu’ elle opère la recommande fortement

son sérieux et sa gentillesse

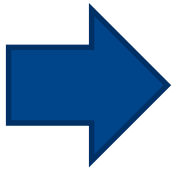


Classe prédite: Infirmière

Vraie classe: Chirurgienne

Détection de biais avec l'attribution

Elle travaille à l'hôpital de Perpignan depuis 3 ans.
Les patients qu' elle opère la recommande fortement
son sérieux et sa gentillesse



Classe prédite: Infirmière

Vraie classe: Chirurgienne

Heatmap de l'importance des mots de l'exemple pour la prédiction de la classe «infirmière»

Détection de biais avec l’attribution

L’enseignante adore aider ses étudiants

The teacher loves to help her

Heatmap de l’importance des mots précédents pour la génération du mot « her »

Détection de biais avec l’attribution

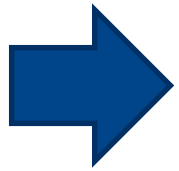
L’enseignante adore aider ses étudiants

The teacher loves to help his

Heatmap de l’importance des mots précédents pour la génération du mot « his »

XAI basée concept pour de la classification

Le Docteur Lecomte travaille au service d'odontologie de l'Hôpital Saint-Martin. Diplômé de la Faculté de Chirurgie Dentaire de Paris, il s'est spécialisé en soins dentaires hospitaliers, notamment dans la prise en charge des patients souffrant de pathologies complexes.

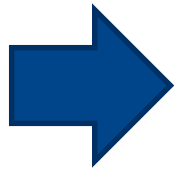


Classe prédite: Dentiste

72 %	Concept 4	*étude supérieure*
63 %	Concept 12	*médicale*
87 %	Concept 32	*dentaire*

XAI basée concept pour de la classification

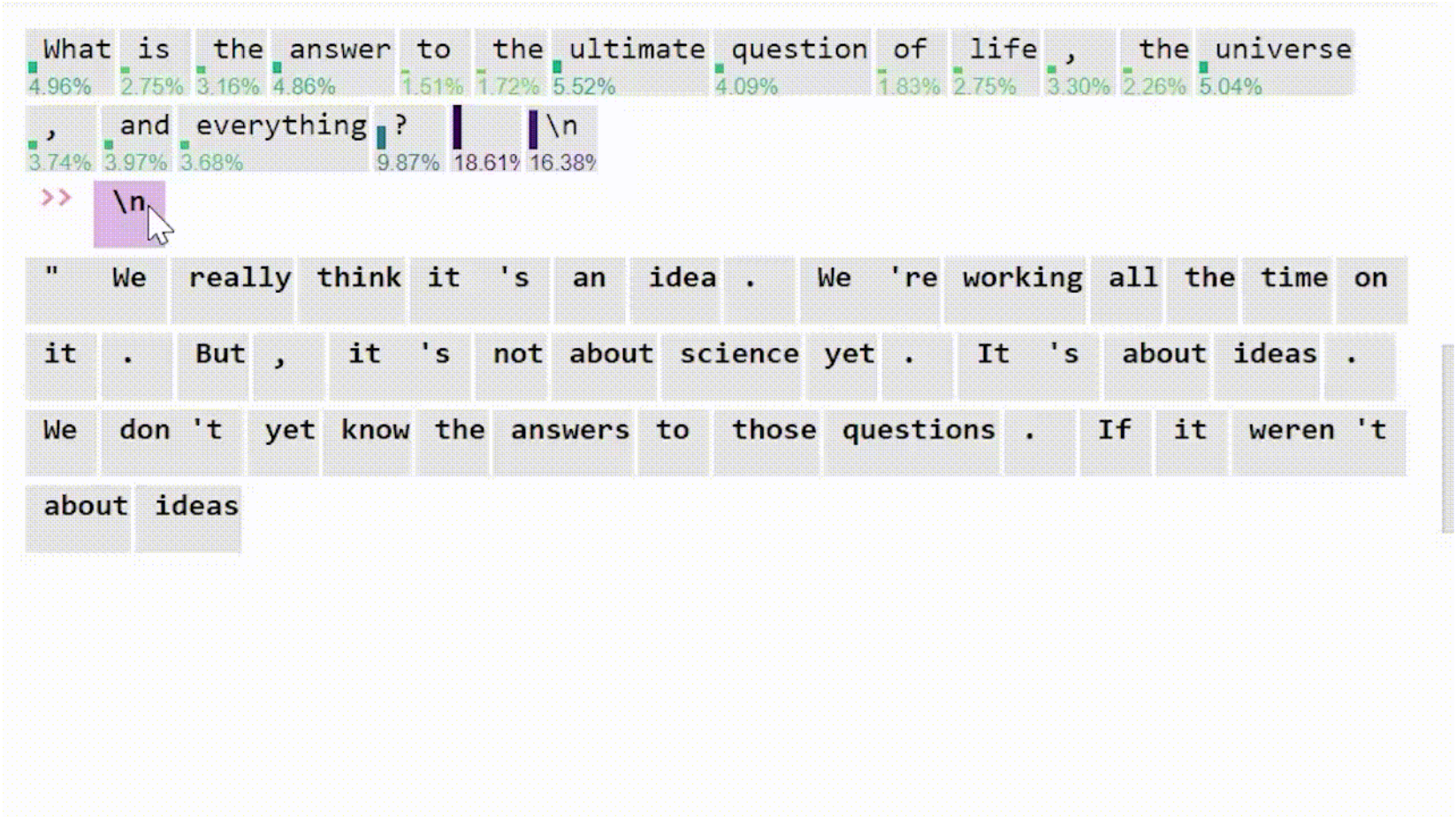
Le Docteur Lecomte travaille au service d'odontologie de l'Hôpital Saint-Martin. Diplômé de la Faculté de Chirurgie Dentaire de Paris, il s'est spécialisé en soins dentaires hospitaliers, notamment dans la prise en charge des patients souffrant de pathologies complexes.



Classe prédite: Dentiste

72 %	Concept 4	*étude supérieure*
63 %	Concept 12	*médicale*
87 %	Concept 32	*dentaire*

XAI basée concept pour de la generation



Source: https://www.eccox.io/

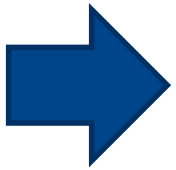
Détection de biais avec de l'XAI basée sur les concepts

DEEL
DEpendable & Explainable Learning

Elle travaille à l'hôpital de Perpignan depuis 3 ans.

Les patients qu' elle opère la recommande fortement

son sérieux et sa gentillesse



Classe prédite: Infirmière

Vraie classe: Chirurgienne

70 % Concept 12 *médicale*

90 % Concept 17 *genre féminin*

Détection de biais avec de l'XAI basée sur les concepts

Elle travaille à l'hôpital de Perpignan depuis 3 ans.
Les patients qu' elle opère la recommande fortement
son sérieux et sa gentillesse

Classe prédite: Infirmière

Vraie classe: Chirurgienne

70 % Concept 12 *médicale*

90 % Concept 17 *genre féminin*

Heatmap de l'importance des mots pour le concept sélectionné pour la prédiction de la classe «infirmière»

DEEL

DEpendable & Explainable Learning

2. Méthodes d'Attribution

a. Basées sur des perturbations

IVADO

IRT SAINT EXUPÉRY

CRIAQ

ANITI Université Fédérale Toulouse Midi-Pyrénées

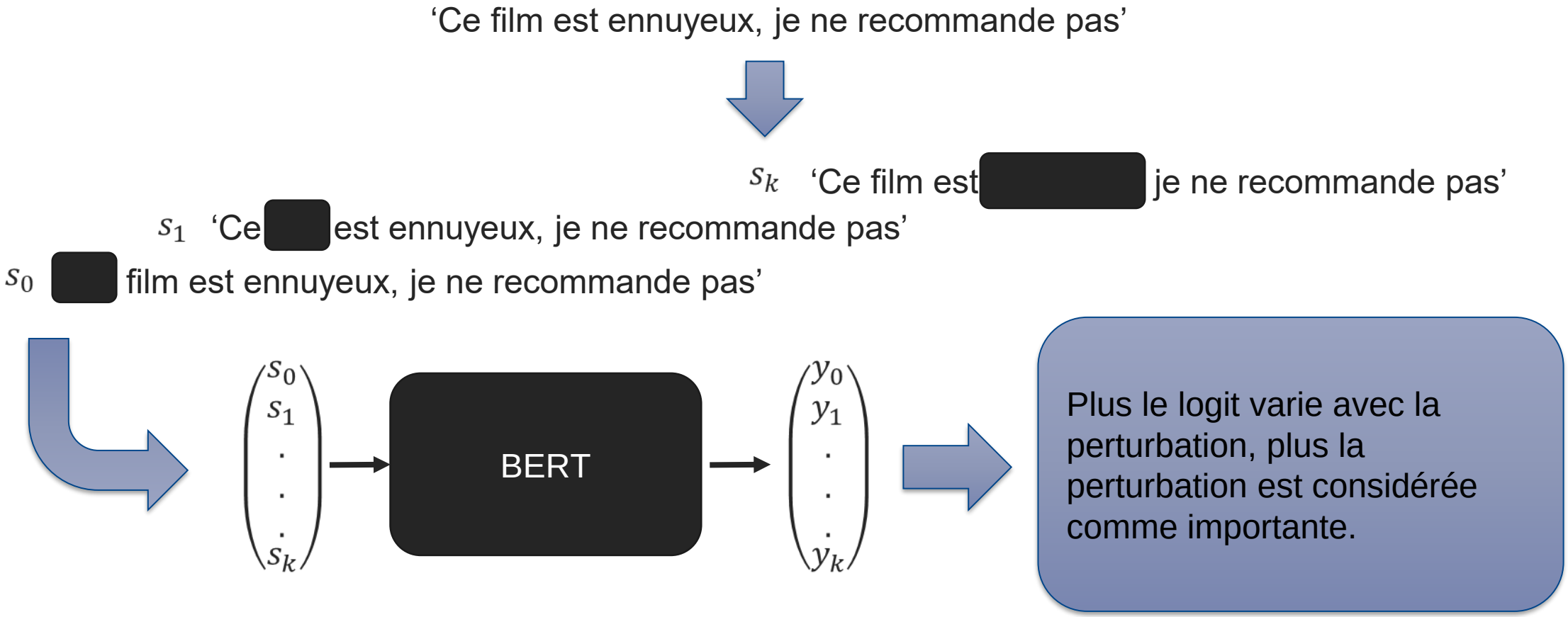
Institut intelligence et données

Attribution basée sur des perturbations

Ces types de méthodes génèrent des explications en perturbant le texte d'entrée : cela peut impliquer de modifier, de supprimer ou de remplacer des mots et d'observer comment ces changements affectent la sortie du modèle.

- Nécessite une exécution répétée du modèle avec des entrées légèrement modifiées.
- Ne requiert pas d’accès aux gradients ou à l’intérieur du modèle (black box).
- Fonctionne avec tout type de modèles (modèle agnostique)

Occlusion



Zeiler and Fergus. "Visualizing and understanding convolutional networks." *Computer Vision–ECCV 2014.*

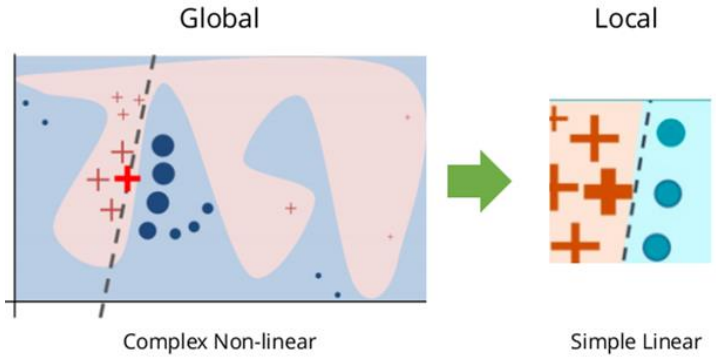
LIME

1. Création d'un jeu de données d'entrainement sur un exemple
Pour une phrase donnée, LIME génère des perturbations en modifiant certains mots.
Chaque phrase perturbée est passée dans le modèle d'origine pour obtenir sa prédiction (score de probabilité).

2. Entraînement du modèle simplifié
Un modèle simple est entraîné en prenant les phrases perturbées comme données d'entrée et les prédictions du modèle d'origine comme cibles.
Chaque mot ou groupe de mots devient une caractéristique (feature) dans le modèle simplifié.

3. Poids extraits
Les coefficients ou poids du modèle simplifié indiquent l'influence de chaque mot sur la prédiction locale.

- Un poids positif signifie que le mot pousse la prédiction vers une certaine classe ou valeur.
- Un poids négatif signifie que le mot éloigne la prédiction de cette classe.



Source: [Interpretability part 3: opening the black box with LIME and SHAP – Kdnuggets]

Source: Ribeiro et al, *"Why should i trust you?" explaining the predictions of any classifier*, ACM SIGKDD 2013.

LIME - Perturbations

Masquage simple :

“Le [MASK] est assis sur le canapé."
“Le chat [MASK] assis sur le canapé."
“Le chat est [MASK] sur le canapé."

Masquage multiple :

“Le [MASK] est [MASK] sur le canapé."
"[MASK] chat [MASK] assis [MASK] le canapé.”

Substitution :

“Le chien est assis sur le canapé."
“Le chat est assis sur le lit."

Source: Ribeiro et al, *"Why should i trust you?" explaining the predictions of any classifier*, ACM SIGKDD 2013.

DEEL

DEpendable & Explainable Learning

2. Méthodes d'Attribution

b. Basées sur les gradients

IVADO

IRT SAINT EXUPÉRY

CRIAQ

ANITI Université Fédérale Toulouse Midi-Pyrénées

Institut intelligence et données

Attribution basée sur les gradients

Utilise les **dérivées partielles** de la sortie du modèle par rapport à ses entrées pour évaluer l'impact de chaque élément d'entrée.

- Dépend des gradients du modèle, donc nécessite un accès direct aux paramètres et aux calculs internes (white box).
- Ne fonctionne qu'avec des modèles différentiables (e.g., réseaux neuronaux).

Saliency map

Les **saliency maps** (cartes de saillance) utilisent les gradients pour évaluer l'importance des éléments d'entrée (e.g., mots, pixels) dans la prédiction d'un modèle.

$$\text{Saillance}(x_i) = \frac{\partial f(x)}{\partial x_i}$$

x_i : Élément d'entrée (e.g., un pixel ou un mot).

$f(x)$: Sortie du modèle pour l'entrée x.

Ces gradients mettent en évidence le token dans le texte d'entrée qui a l'impact le plus significatif sur la prédiction du modèle.

Source: Simonyan et al, *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*, 2013.

Integrated gradients

Integrated Gradients est une technique qui attribue une importance à chaque élément d'entrée en mesurant sa **contribution totale** à la prédiction, en intégrant les gradients le long d'un chemin entre une **référence** et l'entrée réelle.

Source: Sundararajan et al, *Axiomatic Attribution for Deep Networks*, ICML 2017.

Integrated gradients

1. Choisir une référence x′:

Une entrée "neutre" (e.g., une image noire ou une phrase remplie de mot vide) servant de point de départ.

2. Interpoler entre la référence et l'entrée réelle x :

Créer des points intermédiaires $\{x' + \alpha(x - x')\}_{\alpha=0'}^{1}$ où α est un facteur d'échelle.

3. Calculer les gradients le long de ce chemin :

Pour chaque point interpolé, calculer le gradient de la sortie par rapport à l'entrée.

4. Intégrer les gradients :

La contribution d'un élément d'entrée xi est donnée par :

$$IG_i = (x_i - x_i') \int_0^1 \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} d\alpha$$

Source: Sundararajan et al, *Axiomatic Attribution for Deep Networks*, ICML 2017.

DeepLIFT

DeepLIFT est une méthode qui attribue une importance à chaque élément d'entrée en comparant l'effet de cet élément à une **valeur de référence**.

Plutôt que de calculer uniquement les gradients, DeepLIFT analyse **les différences** entre :

- Les activations des neurones pour une entrée donnée x.
- Les activations des neurones pour une entrée de référence x′.

Source: Shrikumar et al, *Learning Important Features Through Propagating Activation Differences*, ICML 2017.

DeepLIFT

1. Choisir une référence x′ (comme pour Integrated gradients)

2. Comparer les activations :
Calculer la différence entre l'entrée x et la référence x′ (pour chaque xi et xi').

3. Rétropropagation des contributions :
Propager les contributions $(\frac{\Delta y}{\Delta x_i})$ dans tout le réseau.

4. Calcule de l'importance pour chaque entrée:
Pour chaque élément d'entrée xi, l'importance est calculée comme :

$$C_{\Delta x_i} = \Delta x_i \cdot \frac{\Delta y}{\Delta x_i}$$

$$\Delta x_i = x_i - x_i'$$
$$\Delta y = f(x) - f(x')$$

$\frac{\Delta y}{\Delta x_i}$: Contribution relative de x_i au changement dans y.

Source: Shrikumar et al, *Learning Important Features Through Propagating Activation Differences*, ICML 2017.

DEEL

DEpendable & Explainable Learning

2. Méthodes d'Attribution

c. Basées sur la structure interne

IVADO

IRT SAINT EXUPÉRY

CRIAQ

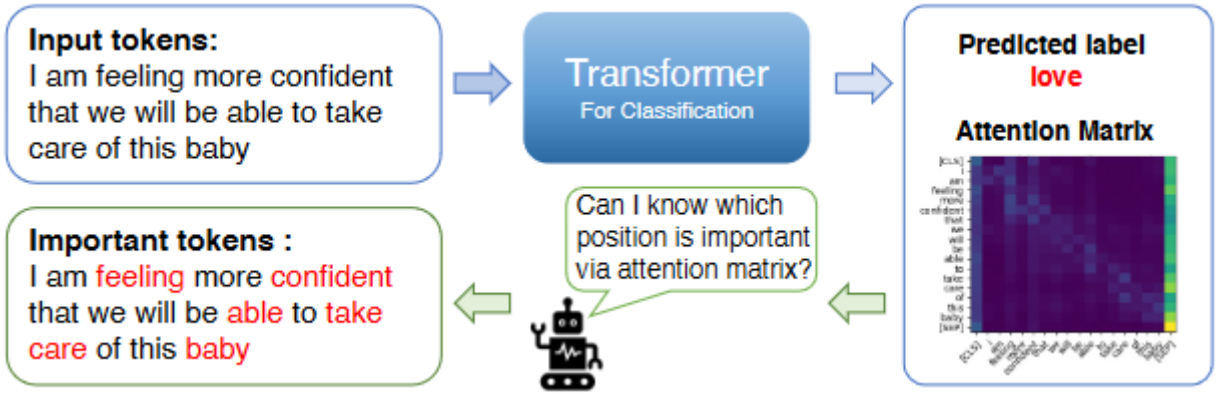
ANITI Université Fédérale Toulouse Midi-Pyrénées

Institut intelligence et données

Attribution avec les poids d'Attention

- Les poids d'Attention (outputs des mécanismes d'Attention dans des modèles comme Transformers) indiquent où le modèle "regarde" lorsqu'il fait une prédiction.

- Ces poids peuvent être utilisés comme des scores d'attribution, attribuant une importance aux éléments d'entrée (e.g., mots, tokens, pixels).



Source: ATTEXPLAINER: Explain Transformer via Attention by Reinforcement Learning, Niu et al. IJCAI 2022.

Attribution avec les poids d'Attention

Avantages :

- **Directement accessible :** Les poids d'attention sont déjà calculés lors de l'inférence, ce qui évite des calculs supplémentaires.
- **Intuitif :** Facile à visualiser pour des tâches comme la traduction ou le résumé

Inconvénients :

- **Complexité multi-couche :** Avec plusieurs têtes et couches d'attention, l'interprétation devient plus difficile.
- **Pas toujours fiable :** Les poids d'attention ne capturent pas nécessairement l'importance causale d'un élément. Ils reflètent où le modèle "regarde," mais pas si cet élément change la prédiction. [1, 2, 3].

[1] Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In NAACL-HLT, 2019.
[2] Sofia Serrano and Noah A. Smith. Is attention interpretable? In ACL, 2019.
[3] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In EMNLP, 2019.

Attribution avec l’Attention Flow

Poids d’attention cumulés :

- Contrairement à l’interprétation directe des poids d’attention d’une seule couche, cette méthode calcule les contributions cumulées en remontant l’information à travers toutes les couches.
- Les poids d’attention sont combinés pour refléter l’impact global d’une caractéristique d’entrée.

Propagation des attributions :

- Les scores d’attention sont propagés à l’envers, des sorties vers les entrées.
- Les contributions sont calculées en suivant la structure du modèle, en tenant compte de toutes les connexions.

Calcul de l’importance :

- L’importance finale d’une entrée est obtenue en agrégeant les contributions à travers toutes les couches.

Source: Abnar and Zuidema, *Quantifying Attention Flow in Transformers*, 2020.

DEEL

DEpendable & Explainable Learning

2. Méthodes d'Attribution

d. Valeurs de Shapley

IVADO

IRT SAINT EXUPÉRY

CRIAQ

ANITI Université Fédérale Toulouse Midi-Pyrénées

Institut intelligence et données

Valeurs de Shapley

Les **valeurs de Shapley** proviennent de la **théorie des jeux coopératifs** et permettent de mesurer la **contribution équitable** de chaque joueur dans un jeu où les participants collaborent pour obtenir un gain collectif.

Pour un ensemble de joueurs N et un joueur i, la valeur de Shapley est donnée par :

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! \cdot (|N| - |S| - 1)!}{|N|!} \cdot [v(S \cup \{i\}) - v(S)]$$

- N : Ensemble total des joueurs.
- S : Sous-ensemble de joueurs sans i.
- v(S) : Gain (ou valeur) obtenu par le sous-ensemble S.
- |S| : Nombre de joueurs dans S.

Source: Shapley, Lloyd S. "Stochastic games." *Proceedings of the national academy of sciences* 39.10 (1953): 1095-1100.

SHAP - SHapley Additive exPlanations

SHAP (SHapley Additive exPlanations) est une méthode générale qui utilise les valeurs de Shapley pour mesurer l'importance de chaque caractéristique (ou "joueur") d'une entrée, afin d'expliquer les prédictions d'un modèle.

Chaque implémentation de SHAP est adaptée à des contextes spécifiques.

Source: Lundberg and Lee, *A Unified Approach to Interpreting Model Predictions*, Neurips 2017.

SHAP méthodes – premières variations

- **KernelSHAP**: Adaptation de LIME - donc agnostique par rapport au modèle - pour approximer les valeurs SHAP. Comme il fonctionne pour n'importe quel modèle f, il ne peut faire aucune hypothèse sur sa structure et est donc le plus lent.

- **DeepSHAP**: Adaptation de DeepLIFT - donc spécifique aux réseaux de neurones - pour approximer les valeurs SHAP. Considérablement plus rapide que KernelSHAP, car il émet des hypothèses sur la nature de la composition du modèle.

- **LinearSHAP**: Spécifique aux modèles linéaires, utilise les coefficients de pondération du modèle et peut prendre en compte les corrélations entre les caractéristiques.

Source: Lundberg and Lee, *A Unified Approach to Interpreting Model Predictions*, Neurips 2017.

SHAP méthodes – d’autres variations

- **PartitionSHAP**: Version plus rapide de KernelSHAP qui regroupe les caractéristiques de manière hiérarchique. Cette hiérarchie définit des coalitions d'éléments sur la base de leurs interactions.

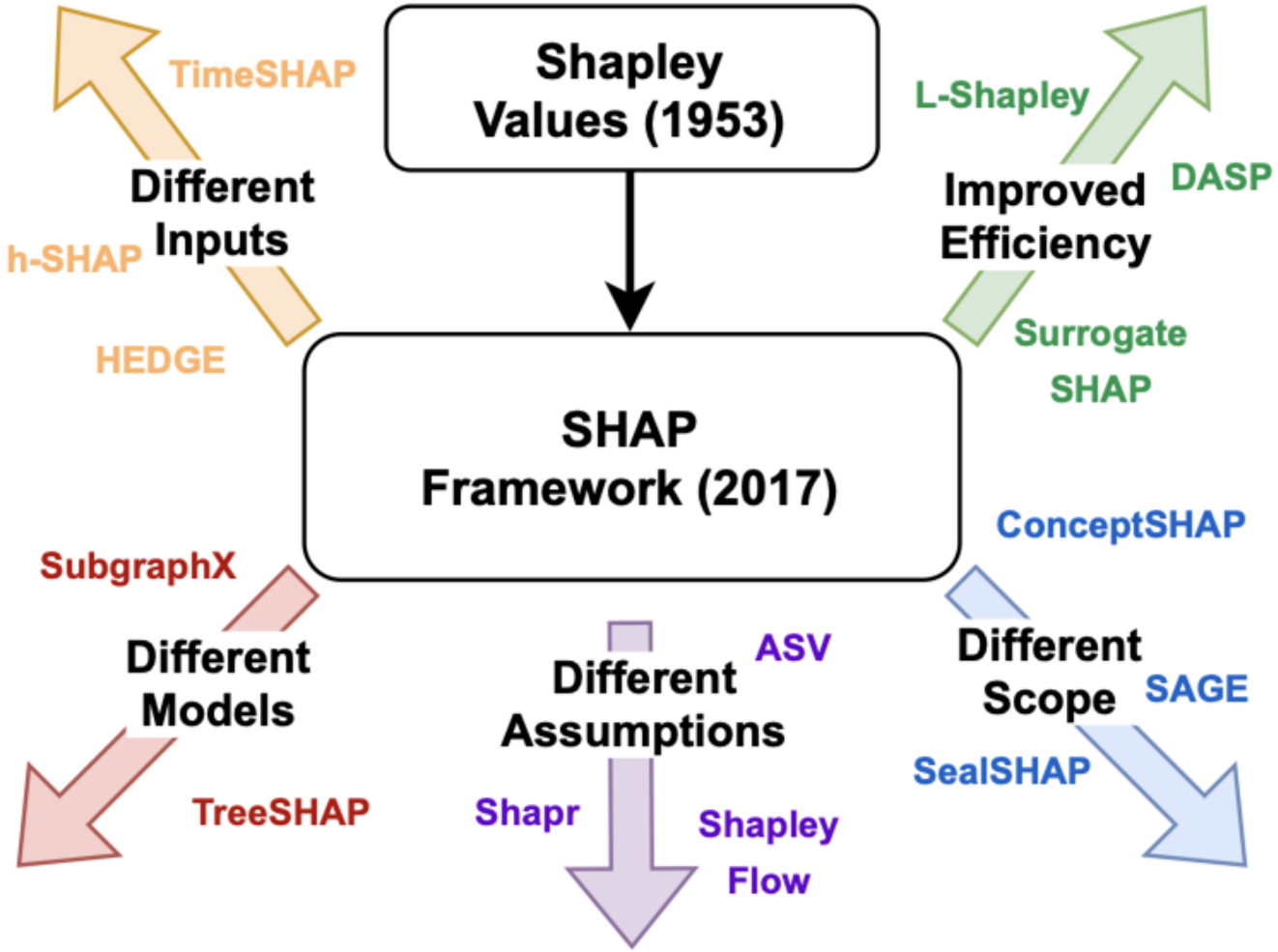
- **GradientSHAP**: Une extension de la méthode *Integrated Gradients* (IG) - spécifique aux réseaux de neurones - qui agrège les gradients sur la différence entre la sortie attendue du modèle et la sortie actuelle.

- **TreeSHAP**: Une méthode rapide pour calculer les valeurs exactes de SHAP pour les arbres et les ensembles. Par rapport à KernelSHAP, elle prend également en compte les interactions entre les caractéristiques.

Source: Mosca et al, *SHAP-Based Explanation Methods: A Review for NLP Interpretability*, Coling 2022.

SHAP-Based Explanation Methods

A Review for NLP Interpretabilites



Source: Mosca et al, *SHAP-Based Explanation Methods: A Review for NLP Interpretability*, Coling 2022.

DEEL

DEpendable & Explainable Learning

3. Méthodes basées sur les concepts

a. Avec ces concepts supervisés

IVADO

IRT SAINT EXUPÉRY

CRIAQ

ANITI Université Fédérale Toulouse Midi-Pyrénées

Institut intelligence et données

Concepts supervisés

Définir des concepts supervisés :

Les concepts sont représentés par des ensembles d'exemples annotés.

Exemple : Le concept "politesse" peut être défini à partir de phrases comme :

''Pourriez-vous m'envoyer le rapport, s'il vous plaît ?'' -> label: ''Polie''

''Donne-moi le rapport tout de suite !'' -> label: ''Impolie''

Dans le cas d'une étude pour la Fairness, la variable « genre » pourrait devenir un concept.

Testing with Concept Activation Vector (TCAV)

Approche Post-hoc sur un modèle déjà entrainé.

1. Définir des concepts supervisés

2. Projeter dans l'espace conceptuel :

- Pour créer le vecteur concept, on agrège toutes les représentations internes du modèle (e.g., les embeddings) correspondantes aux exemples du concept.
- Les embeddings sont projetées pour mesurer la présence ou l'importance des concepts.

3. Attribuer l'importance des concepts :

Des techniques évaluent comment les concepts influencent la prédiction finale.

Source: Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." ICML 2018.

Concept Bottleneck Models (CBM)

Entrainement d'un modèle plus explicable.

1. Définir des concepts supervisés

2. Création du modèle en deux parties:

- La première partie prédit les concepts supervisés depuis l'entrée.
- Une seconde partie utilise uniquement ces concepts pour prédire la sortie finale.

Les explications sont directement intégrées au modèle.

Source: Koh, Pang Wei, et al. "Concept bottleneck models." ICML 2020.

DEEL

DEpendable & Explainable Learning

3. Méthodes basées sur des concepts

b. Avec des concepts non-supervisés

IVADO

IRT SAINT EXUPÉRY

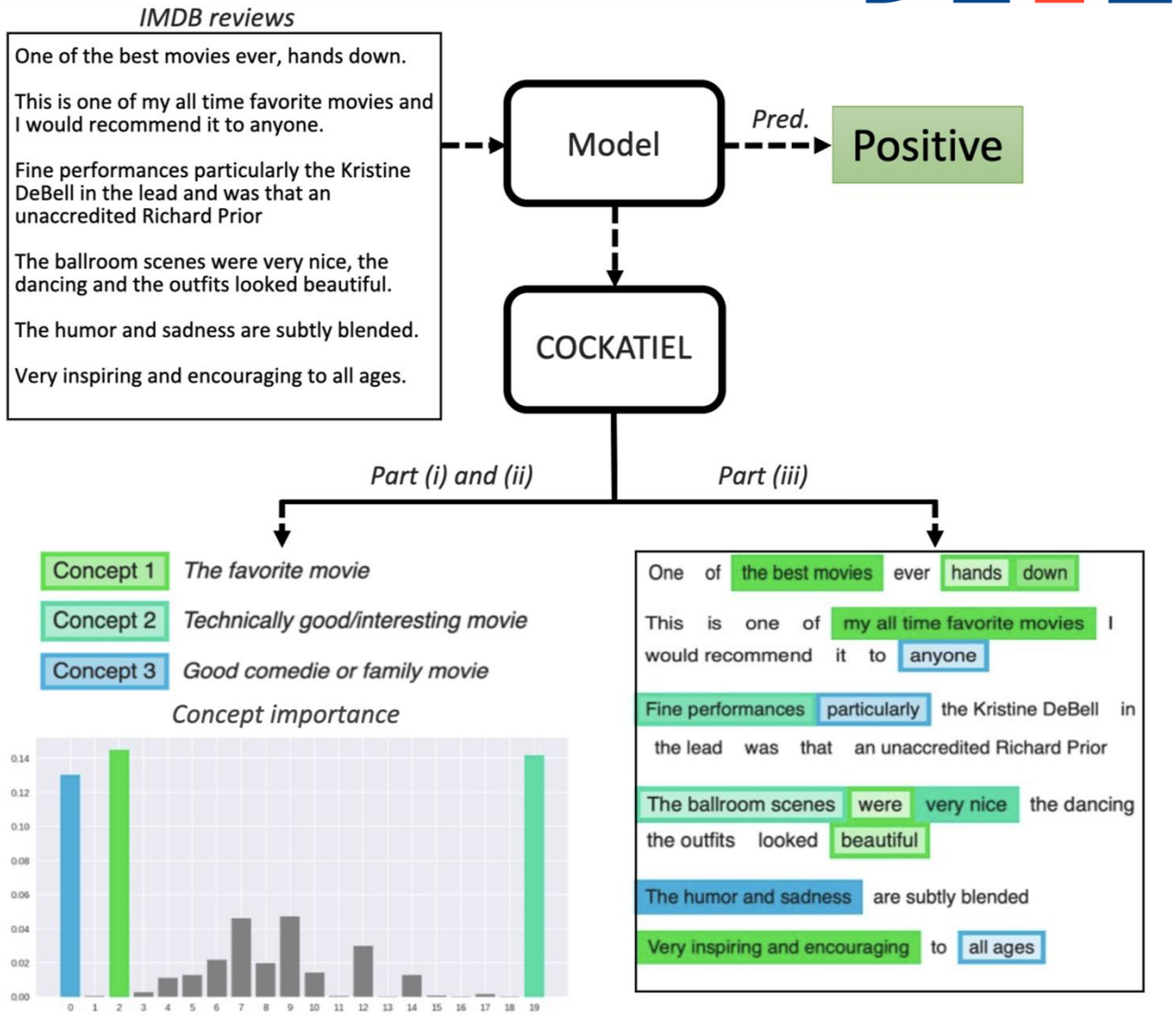
CRIAQ

ANITI Université Fédérale Toulouse Midi-Pyrénées

Institut intelligence et données

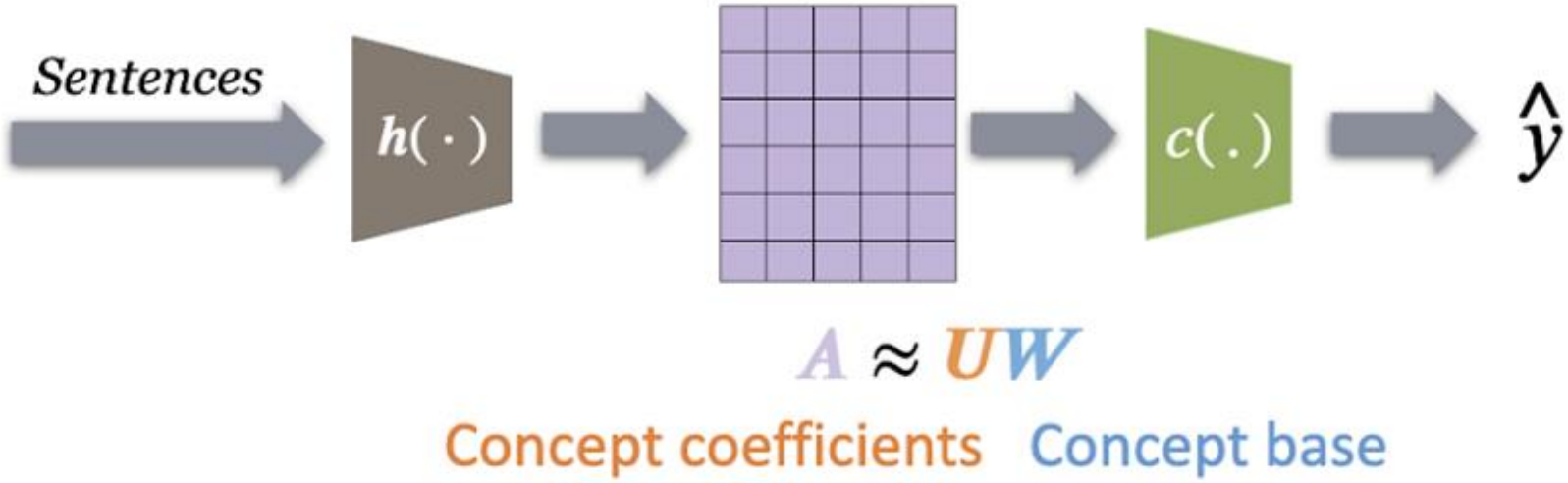
COCKATIEL

- Première méthode basée sur des concepts non supervisés pour le NLP.
- Méthode modèle agnostique mais nécessité d'avoir accès à la dernière couche du modèle (pas totalement blackbox).



Source: Jourdan et al, *COCKATIEL: COntinuous Concept ranKed ATtribution with Interpretable ELements for explaining neural net classifiers on NLP tasks*, ACL 2023.

COCKATIEL
1) Découverte non supervisée de concepts



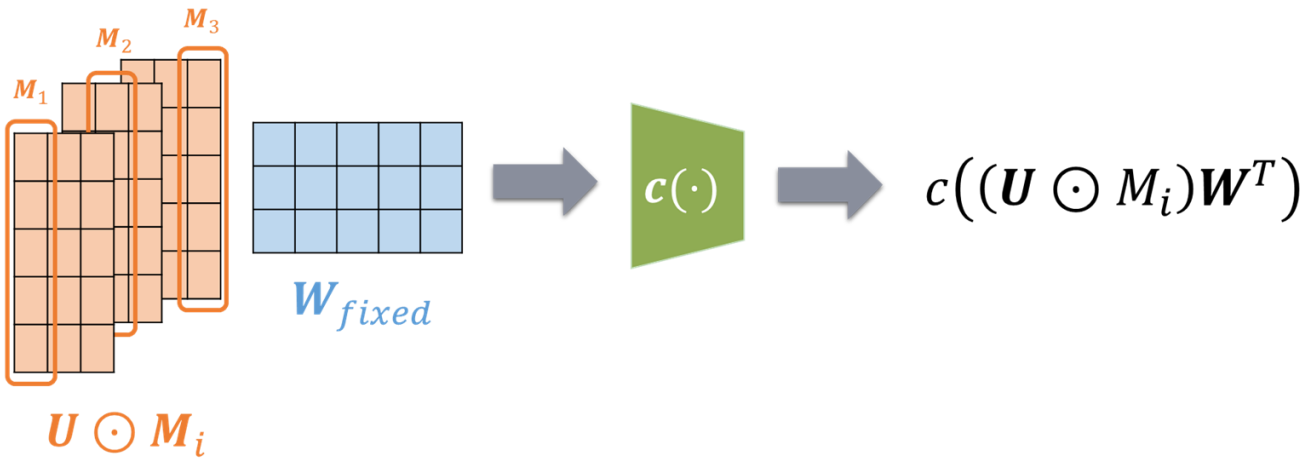
Décomposition de l'embedding final du modèle (Token CLS) en deux matrices : le coefficient de concept et la base de concept.

Décomposition NMF pour découvrir des concepts pour chaque classe prédite.

Source: Jourdan et al, *COCKATIEL: COntinuous Concept ranKed ATtribution with Interpretable ELements for explaining neural net classifiers on NLP tasks*, ACL 2023.

COCKATIEL
2) Estimation de l'importance des concepts

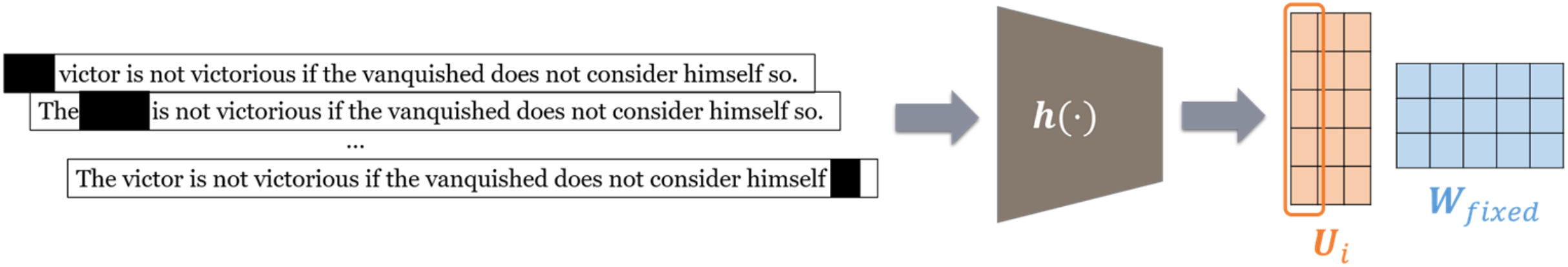
- La matrice U est perturbée pour observer l'effet de ces perturbations sur les résultats du modèle.
- Plus la variance de la sortie est grande lors de la perturbation d'un concept donné, plus celui-ci est important pour le modèle.



$$\mathcal{S}_{T_i} = \frac{\mathbb{E}_{M_{\sim i}}[\mathbb{V}_{M_i}[c((\boldsymbol{U} \odot \boldsymbol{M})\boldsymbol{W}^T)|\boldsymbol{M}_{\sim i}]}{\mathbb{V}[(\boldsymbol{U} \odot \boldsymbol{M})\boldsymbol{W}^T]}$$

Source: Jourdan et al, *COCKATIEL: COntinuous Concept ranKed ATtribution with Interpretable ELements for explaining neural net classifiers on NLP tasks*, ACL 2023.

COCKATIEL
3) Instance level Explanation Generation



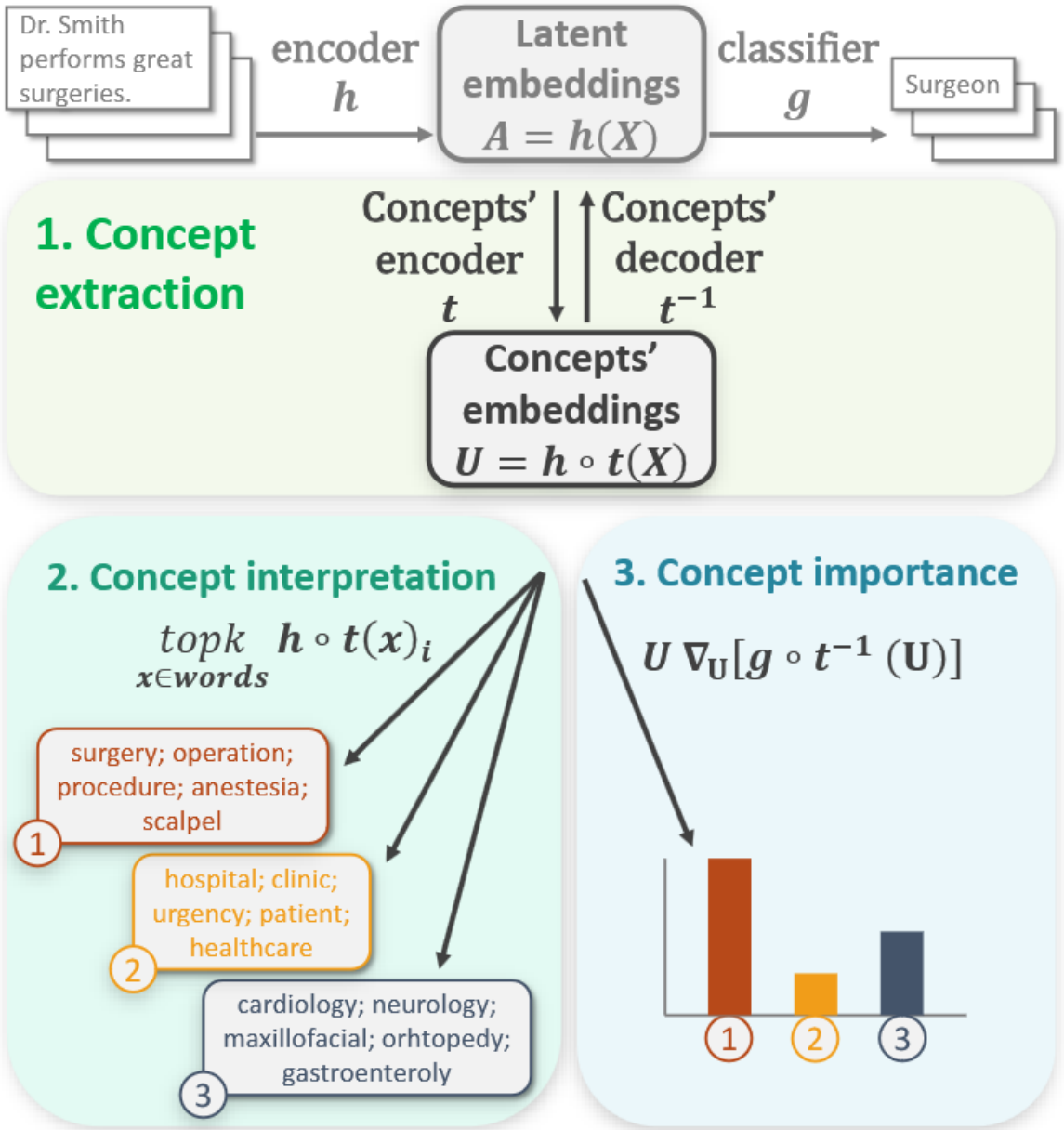
Un mot/clause est retiré du texte d'entrée étudié. Plus il affecte le coefficient correspondant à un concept donné, plus le mot/la clause est important pour le concept.

Source: Jourdan et al, *COCKATIEL: COntinuous Concept ranKed ATtribution with Interpretable ELements for explaining neural net classifiers on NLP tasks*, ACL 2023.

Méthodes avec des concepts non supervisés

Formalisation générale des approches basées sur des concepts non supervisés.

1. L'extraction de l'espace concept U peut être vu comme une fonction t.
2. On peut interpréter ces vecteurs de différentes manières.
3. L'importance des concepts peut être calculé avec toutes les méthodes d'attribution vu précédemment.



Source: Poché, Antonin, et al. "ConSim: Measuring Concept-Based Explanations' Effectiveness with Automated Simulatability." *arXiv preprint* 2025.

Utiliser l’explicabilité basée concept pour réduire les biais

DEEL
DEpendable & Explainable Learning

Supprimer l’information sensible (sous forme de concept) de manière linéaire:

Ravfogel et al. "Null it out: Guarding protected attributes by iterative nullspace projection." *ACL* (2020)

Ravfogel et al. "Linear adversarial concept erasure." *ICML*, 2022

Belrose et al. "Leace: Perfect linear concept erasure in closed form." *Advances in Neural Information Processing Systems* 36 (2024).

Supprimer l’information sensible (sous forme de concept) de manière non-linéaire:

Jourdan et al, “TaCo: Targeted Concept Erasure Prevents Non-Linear Classifiers From Detecting Protected Attributes”, *arxiv* 2024

DEEL

DEpendable & Explainable Learning

4. Explication Contrastive et Contrefactuelle

IVADO

IRT SAINT EXUPÉRY

CRIAQ

ANITI Université Fédérale Toulouse Midi-Pyrénées

Institut intelligence et données

Contrastive vs Contrefactuelle

DEEL
DEpendable & Explainable Learning

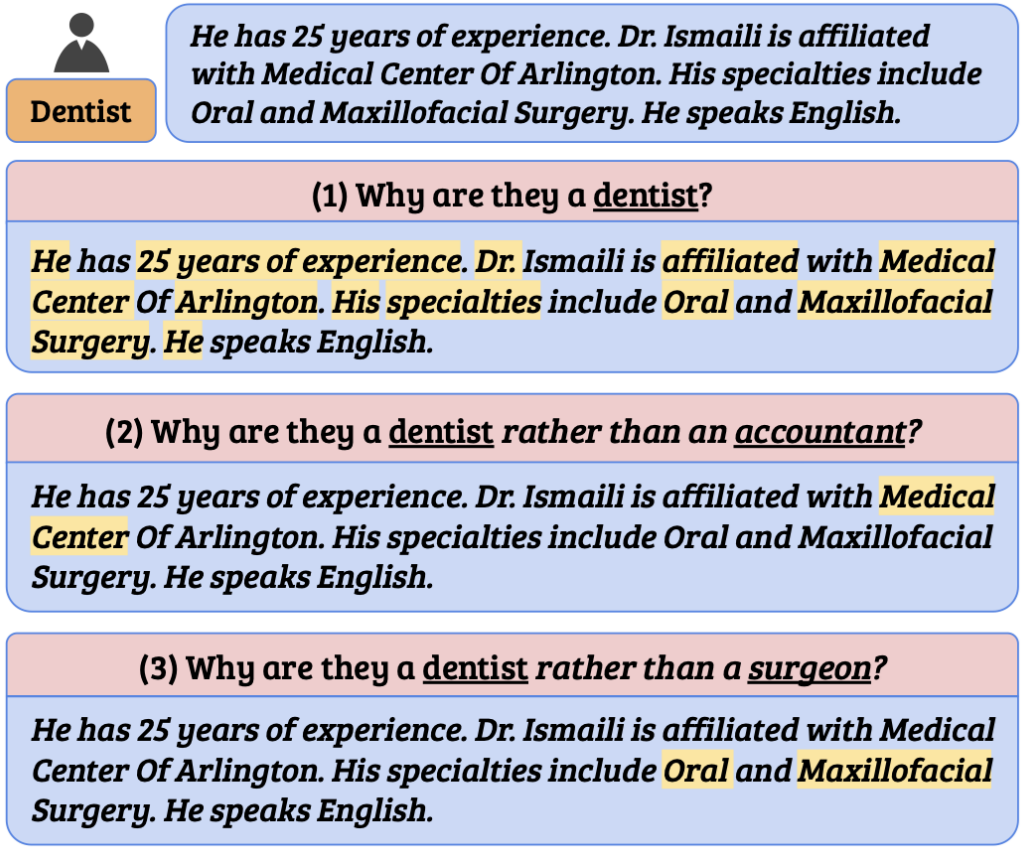
Contrastive

- Se concentre sur les raisons pour lesquelles le modèle n'a pas choisi un autre résultat spécifique qui aurait pu être attendu ou typique.
- Utile dans les situations où les utilisateurs ont besoin d'explications sur le processus de prise de décision par rapport à une alternative particulière.

Contrefactuelle

- Exploration de ce qui pourrait être modifié dans les données d'entrée pour obtenir un résultat différent, en se concentrant sur les changements minimaux qui conduisent à une modification des résultats.
- Utile pour améliorer les modèles en testant leurs réponses à des changements nuancés, ce qui peut aider à identifier les biais du modèle.

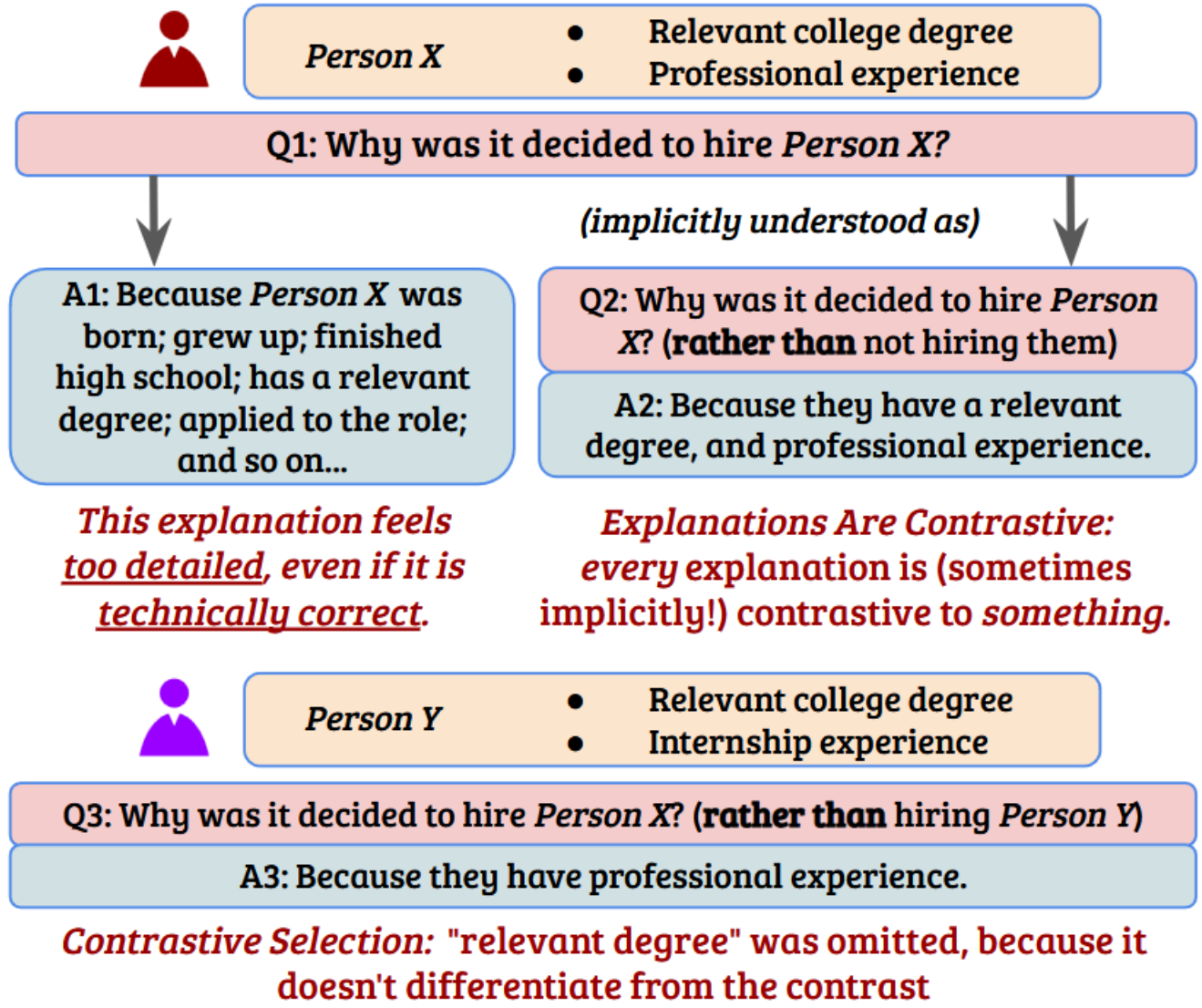
Explication contrastive pour l'attribution



Les explications sans contraste explicite (1) sont potentiellement mal alignées sur les attentes humaines, ce qui les rend confuses pour l'interprétabilité humaine.

Les explications contrastives (2, 3) réduisent l'espace de tous les facteurs de causalité à ceux qui sont « intuitivement » pertinents, facilitant ainsi une compréhension plus fine, et peuvent varier en fonction de la décision de contraste (par exemple, comptable, chirurgien).

Source: Jacovi et al, *Contrastive Explanations for Model Interpretability,* EMNLP 2021.



Source: Jacovi et al, *Contrastive Explanations for Model Interpretability*, EMNLP 2021.

Explication contrastive pour la génération

Question:

Ann and her children are going to Linda's home ___.
(a) by bus (b) by car (c) on foot (d) by train

Original Context:

...Dear Ann, I hope that you and your children will be here in two weeks. My husband and I will go to meet you at the train station. Our town is small...

Prediction: (d) *by train*

Why *by train* (d) and not *on foot* (c)?

MiCE-Edited Context:

...Dear Ann, I hope that you and your children will be here in two weeks. My husband and I will go to meet you at ~~the train station~~ **your home on foot**. Our ~~town~~ **house** is small...

Contrast Prediction: (c) *on foot*

MICE génère des explications contrastives sous la forme de modifications des entrées qui changent les prédictions du modèle en prédictions de la cible (contraste).

Dans l'exemple, l'édition (en gras et en rouge) est minimale et fluide, et elle modifie la prédiction du modèle de « *by train* » à « *on foot* » (surlignée en gris).

Source: Ross et al, *Explaining NLP Models via Minimal Contrastive Editing (MICE)*, Findings of ACL 2021.

Méthodes contrefactuelles

Control code	Definitions and POLYJUICE-generated Examples	Training Datasets
`negation`	A dog is not embraced by the woman.	(Kaushik et al., 2020)
`quantifier`	A dog is → Three dogs are embraced by the woman.	(Gardner et al., 2020)
`shuffle`	*To move (or swap) key phrases or entities around the sentence.* A dog → woman is embraced by the woman → dog.	(Zhang et al., 2019b)
`lexical`	*To change just one word or noun chunk without altering the POS tags.* A dog is embraced → attacked by the woman.	(Sakaguchi et al., 2020)
`resemantic`	*To replace short phrases without altering the remaining dependency tree.* A dog is embraced by the woman → wrapped in a blanket.	(Wieting and Gimpel, 2018)
`insert`	*To add short phrases without altering the remaining dependency tree.* A dog is embraced by the little woman.	(McCoy et al., 2019)
`delete`	*To remove short phrases without altering the remaining dependency tree.* A dog is embraced by the woman.	(McCoy et al., 2019)
`restructure`	*To alter the dependency tree structure,* e.g., *changing from passive to active.* A dog is embraced by → hugging the woman.	(Wieting and Gimpel, 2018)

Source: Wu et al, *Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models*, ACL 2021.

anny.jourdan@irt-saintexupery.com

DEEL
France-Quebec

Merci pour votre attention !

IVADO
IRT SAINT EXUPÉRY
CRIAQ
ANITI Université Fédérale Toulouse Midi-Pyrénées
Artificial & Natural Intelligence Toulouse Institute
Institut intelligence et données