



Interprétabilité A Priori et Explicabilité A Posteriori dans le Traitement Automatique des Langues

Tom BOURGEADE - IRIT - MELODI
tom.bourgeade@irit.fr



Contexte :

Des problèmes de transparence en TAL

Évolution de la complexité des architectures de l'État de l'Art en TAL
(M = Million ; B = Milliard ; T = Trillion)

Model	Parameters	Depth
INFERSENT [Conneau et al., 2017]	~ 50M	4
ELMo [Peters et al., 2018]	~ 100M	4
GPT [Radford et al., 2018]	~ 117M	24
BERT [Devlin et al., 2019]	~ 336M	24
GPT-2 [Radford et al., 2019]	~ 1.5B	48
GPT-3 [Brown et al., 2020]	~ 175B	96

Switch Transformers [Fedus et al., 2021] ~1.6T*

PaLM [Chowdhery et al., 2022] ~540B



Contexte : Des problèmes de données en TAL

Un exemple type : **SNLI** (Stanford Natural Language Inference) [\[Bowman et al., 2015\]](#)

Détection d'inférences textuelles (*NLI : Natural Language Inference*)

↪ “Est-ce que l'hypothèse peut être déduite de la prémisse ?”

Premise	Label	Hypothesis
An older and younger man smiling.	neutral	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment	Some men are playing a sport.



Contexte : Des problèmes de données en TAL

Un exemple type : **SNLI** (Stanford Natural Language Inference) [\[Bowman et al., 2015\]](#)

Détection d'inférences textuelles (*NLI : Natural Language Inference*)

↪ “Est-ce que l'hypothèse peut être déduite de la prémisse ?”

Premise	Label	Hypothesis
[REDACTED]	neutral	Two men are smiling and laughing at the cats playing on the floor.
[REDACTED]	contradiction	A man is driving down a lonely road.
[REDACTED]	entailment	Some men are playing a sport.

Modèles hypothèses-seules $\Rightarrow$ ~ 67 % de précision (jeu de test) $\approx 2 \times$ Précision Aléatoire
[\[Gururangan et al., 2018; Poliak et al., 2018\]](#)



Contexte : Biais de distributions et corrélations indésirables

Termes “Raccourcis” : Corrélation Terme $\leftrightarrow$ Classe

- **Faibles occurrences** individuelles (~100 vs. 500k instances) mais **fortes corrélations**.
- **Apprentissage par Raccourcis** :
 - Biais facile à apprendre $\rightarrow$ ~~Généralisation~~
 - Pas nécessairement détectable à l'évaluation
- **Causes** : “Facilités” d’annotation $\Leftarrow$ Myriadisation (*crowdsourcing*)

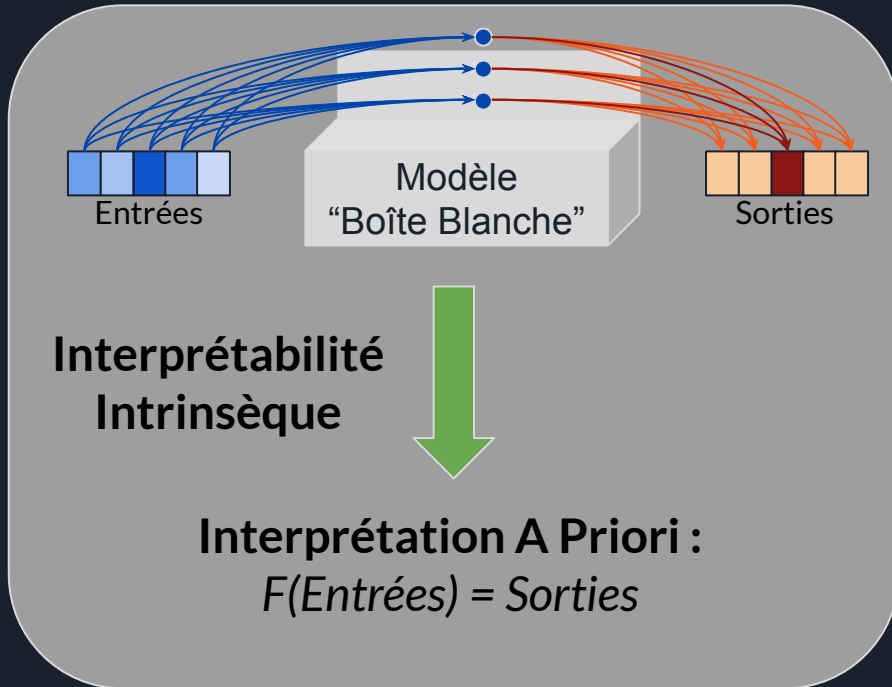
<i>Contradiction</i>		
Word	Prob	Occ
sleeping	0.88	108
driving	0.81	53
Nobody	1.00	52
alone	0.90	50
cat	0.84	49
asleep	0.91	43
no	0.84	31



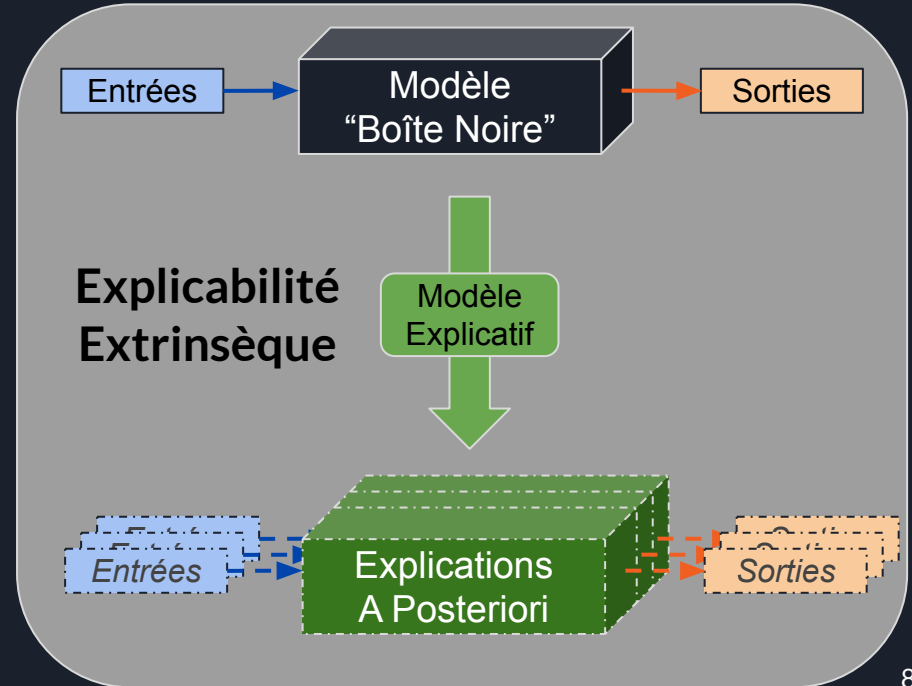
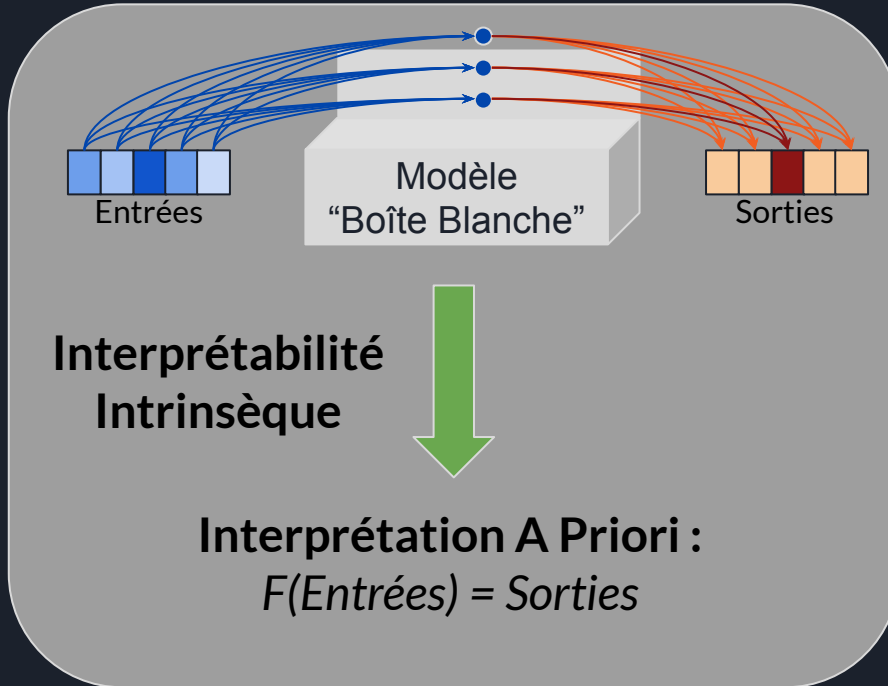
Contexte : Explications

- Expliquer une ou plusieurs **décisions algorithmiques**
- **Pourquoi ?**
 - Quelles **caractéristiques des données ?**
 - Quels **raisonnements internes/règles du modèle ?**  **décisions algorithmiques**
- **Compréhensibilité** (Simplicité, Brièveté, Spécificité) $\leftrightarrow$ **Fidélité** (Précision)
- Explications $\leftrightarrow$ Modèles “Boîtes Noires”

Contexte : Interprétabilité et Explicabilité



Contexte : Interprétabilité et Explicabilité





Plan

- Interprétabilité en TAL :
 - Contexte : Plongements de Mots Interprétables
 - Contributions : Détection de Biais Cachés
- Explicabilité en TAL :
 - Contexte : Méthodes d'Explicabilité
 - Contributions : Génération d'Explications Naturelles

Interprétabilité en TAL - Contexte :

Plongements de Mots Denses

Plongements Denses

One-hot encoding

	cat	mat	on	sat	the
the =>	0	0	0	0	1
cat =>	1	0	0	0	0
sat =>	0	0	0	1	0
...					



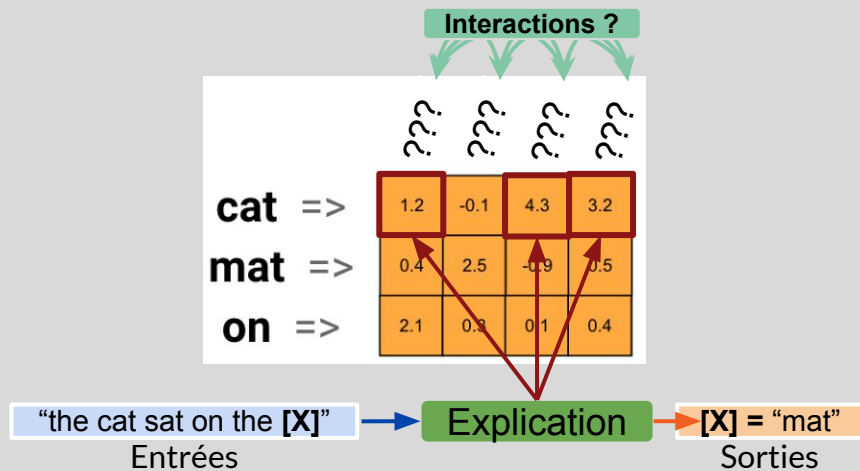
Vecteurs de Plongement de Mots

cat =>	1.2	-0.1	4.3	3.2
mat =>	0.4	2.5	-0.9	0.5
on =>	2.1	0.3	0.1	0.4

Interprétabilité en TAL - Contexte :

Plongements de Mots Denses - Désavantages

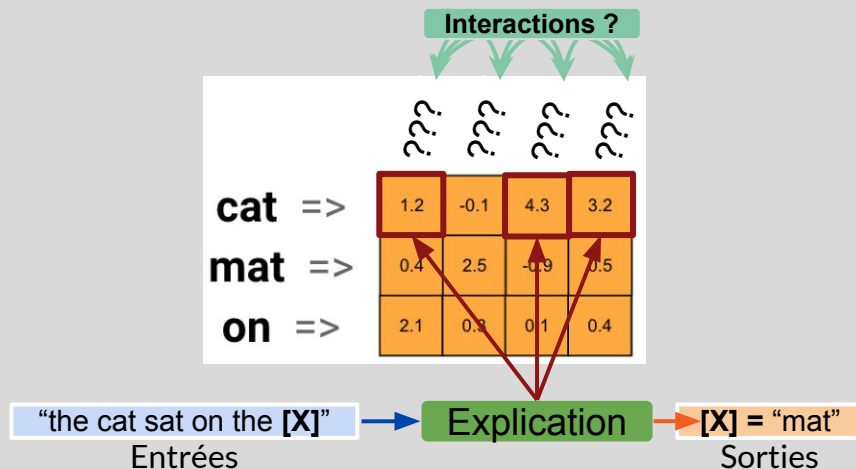
Plongements Denses



Interprétabilité en TAL - Contexte :

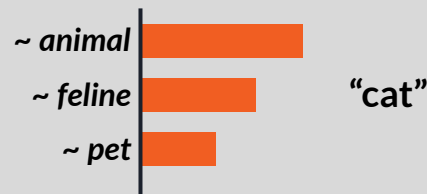
Plongements Parcimonieux et Non-Négatifs

Plongements Denses



Plongements Interprétables

- Non-négativité + Parcimonie + (Normalisation)
→ Distribution de “sens/thématiques”



- Dimensions interprétables :
caractérisation → mots les plus “actifs”

⇒ **Factorisation Non-Négative** de Matrices
[Lee and Seung, [1996](#); [1999](#); [2001](#)]

Interprétabilité en TAL - Contributions :

Modèle de Plongement Interprétable Simple - NMF300

- Factorisation Non-Négative → Statistiques de Cooccurrences termes-termes (Wikipédia)

Mot
"Cible"

N mots les plus actifs dans les 3 dimensions les plus actives du mot "cible"

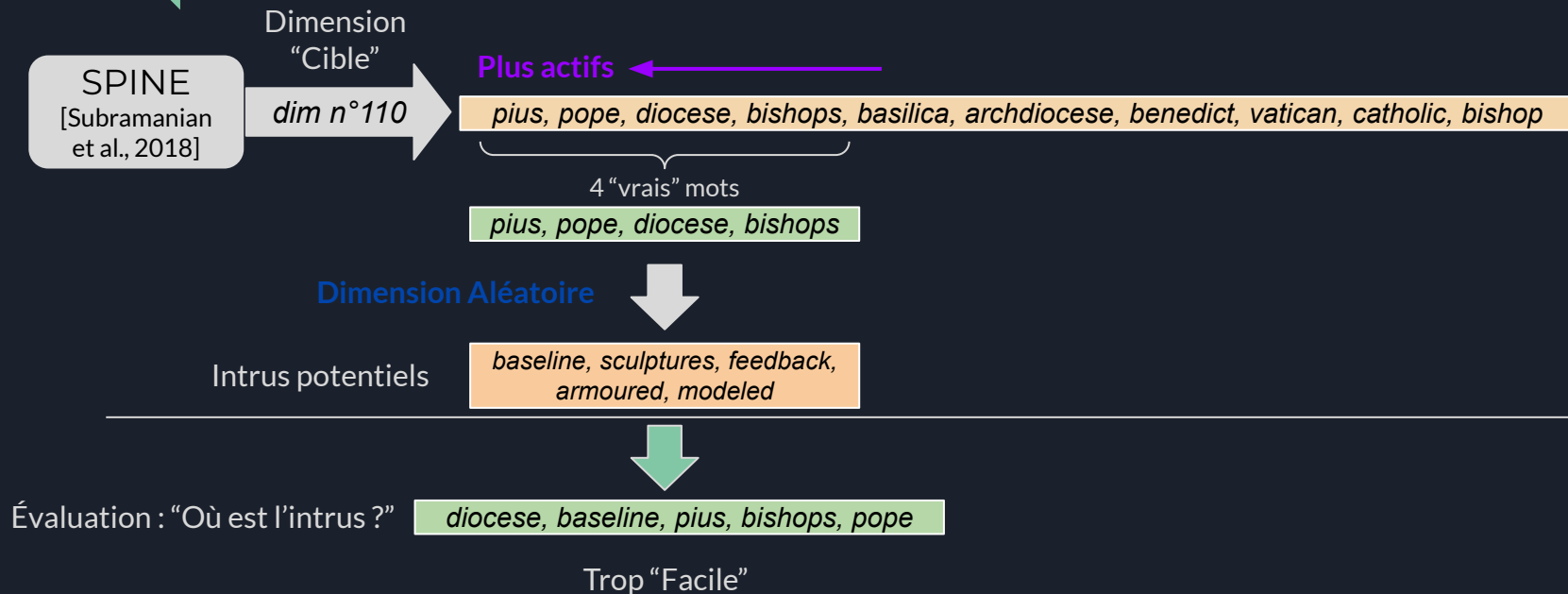
Interprétable		NMF300
science	neurology, ophthalmology, oncology, radiology, microbiology, cardiology	↑
	courses, curriculum, undergraduate, students, vocational, teaching	
	harvard, phd, doctorate, caltech, yale, swarthmore	
Dense		word2vec
science	insult, rivalries, reactors, mw, armistice, massacre	↑
	editing, airplay, cds, professionally, songwriter, screenplay	
	ss, rbi, viii, 2d, xiii, shortstop	

- 4 autres modèles de la littérature : NNSE [\[Murphy et al., 2012\]](#) ; SPOWV [\[Faruqui et al., 2015\]](#) ; SPINE [\[Subramanian et al., 2018\]](#) ; Word2Sense [\[Panigrahi et al., 2019\]](#)

Interprétabilité en TAL - Contributions :

Évaluation - Détection d'Intrusion de Mots

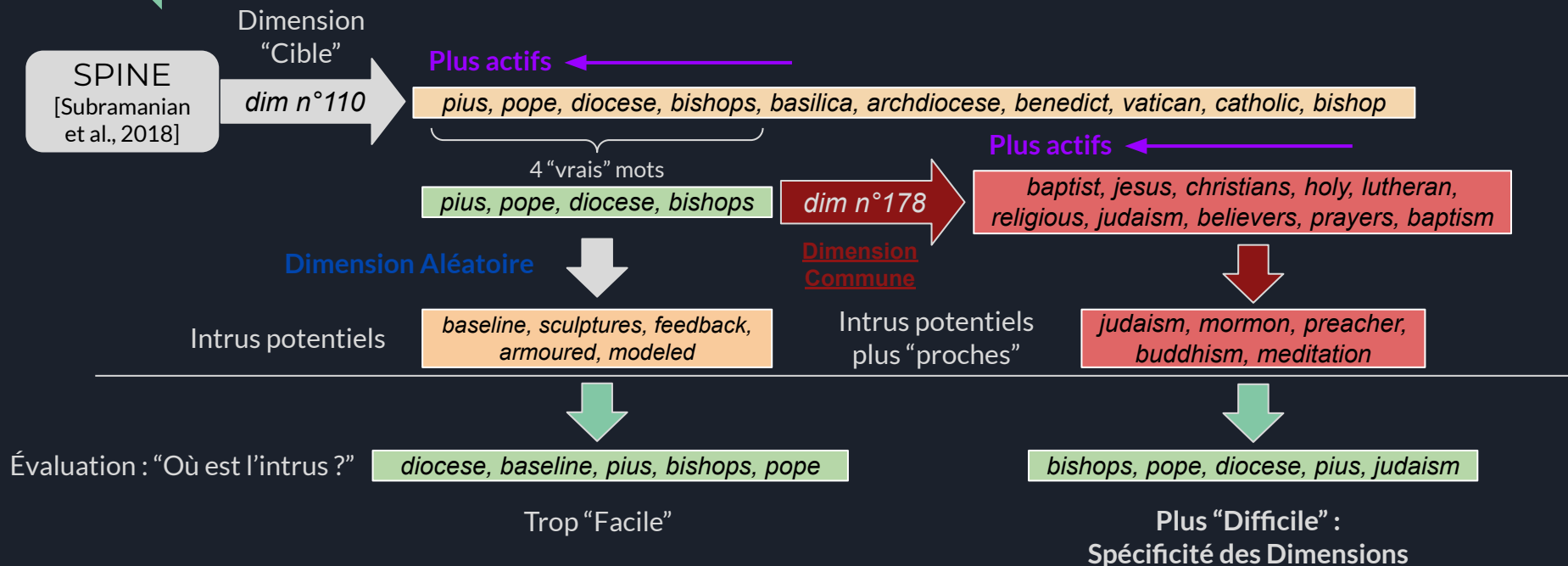
[Chang et al., 2009]



Interprétabilité en TAL - Contributions :

Évaluation - Détection d'Intrusion de Mots

[Chang et al., 2009]



Interprétabilité en TAL - Contributions :

Évaluation - Détection d'Intrusion de Mots - Résultats

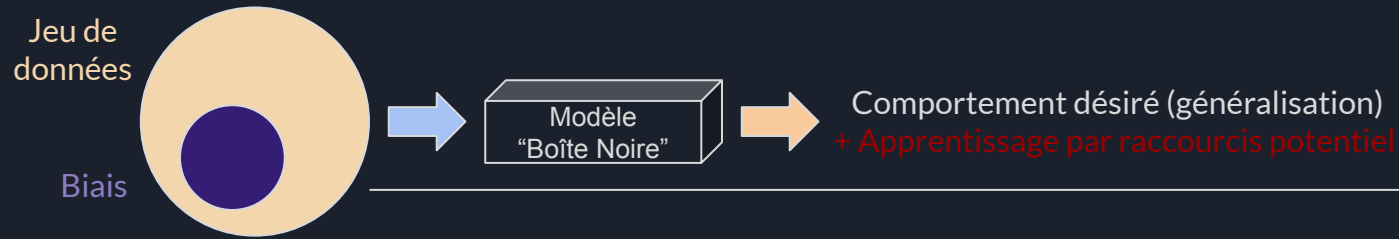
Model	Average Evaluator Accuracy	Inter-evaluator Agreement	Fleiss' Kappa
NMF300	76%	94% ; 72%	0.74
NNSE	79%	90%; 74%	0.76
SPOWV	38%	84%; 34%	0.43
SPINE	79%	92%; 60%	0.63
WORD2SENSE	65%	88%; 56%	0.61

- Résultats avec Plongements Denses $\approx$ Précision Aléatoire

Interprétabilité en TAL - Contributions :

[He et al., 2019]

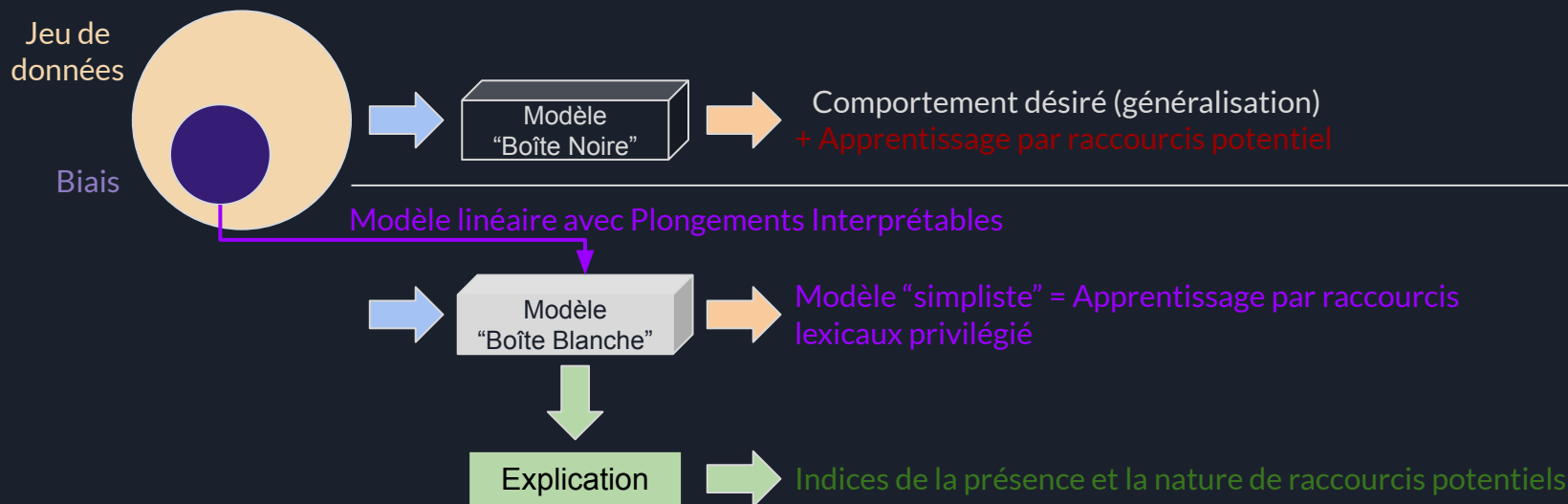
Détection de Biais de Données



Interprétabilité en TAL - Contributions :

[He et al., 2019]

Détection de Biais de Données



Interprétabilité en TAL - Contributions :

Évaluation - Tâches en Aval

Jeu de Données	Tâche	Nombre de Classes	Exemple	Label
IMDB [Maas et al., 2011]	Analyse de Sentiment	2	<i>"This film is just plain horrible[...]"</i>	Négatif
BoolQ [Clark et al., 2019]	Question-Réponses (Vrai/Faux)	2	<i>"do Iran and Afghanistan speak the same language?"</i> + [article associé]	Vrai
Sarcasm [Oraby et al., 2016]	Détection du Sarcasme	2	(1) <i>"americans as a whole don't say soccer sucks; we just don't give a rat's xxx about it. "</i> (2) <i>"the same way most people on planet earth dont give a xxxx about american "football" "</i>	Sarcastique
SNLI [Bowman et al., 2015]	Détection d'Inférences Textuelles	3	(1) <i>"Children smiling and waving at camera"</i> (2) <i>"There are children present"</i>	Entailment
PDTB [Prasad et al., 2008]	Classification de relations rhétoriques	11	(1) <i>"In July, the EPA banned asbestos"</i> (2) <i>"By 1997, all uses of asbestos will be outlawed"</i>	Result

Interprétabilité en TAL

Évaluation - Tâches en Aval

Model \ Corpus	IMDB	BOOLQ	SARCASM	UR-FUNNY	SST	SNLI	EMERGENT	PDTB
NMF300	67.8	62.6	60.5	57.7	54.6	58.6	50.9	33.2
NNSE	78.7	63.6	63.9	59.9	60.6	56.3	66.8	31.1
SPOWV	81.9	66.9	70.5	65.0	62.9	62.9	72.2	36.6
SPINE	81.3	65.9	67.8	63.6	59.9	64.1	72.2	34.5
WORD2SENSE	82.2	66.2	67.3	63.9	61.4	65.5	69.8	34.2
DUMMY (baseline)	50.5	53.5	53.0	52.5	39.5	33.6	41.3	19.3
FASTTEXT	82.0	63.7	70.1	64.5	64.4	61.3	69.5	33.4
Dedicated models*	96.8	76.9	74 [†]	64.4	96	91.5	73	48

- Résultats surprenamment élevés (≡ Modèles dédiés) :
 - Composante purement lexicale importante dans de nombreux jeux de données
 - Problème potentiel (Apprentissage par raccourcis : $\sum(\text{termes "positifs"}) \neq \text{sentiment "positif"}$)
"This movie was not good! I really did not enjoy it! I would not gladly recommend it to all my friends!"
[[Naik et al., 2018](#); [Kassner and Schütze, 2020](#); [Hossain et al., 2020](#)]
- Résultats hétérogènes selon le Modèle de Plongement (NMF300 : certaines classes jamais prédites)
- Tâches "difficiles" : PDTB (connu) et SNLI (raccourcis plus structuraux que lexicaux, e.g. négation)

Interprétabilité en TAL - Contributions :

Évaluation - Diagnostics de Corpus

Dataset	Model	Class	h	C	Most active words in h -th dimension
IMDB	NNSE	<i>pos</i>	192	1.0	<i>utmost, sheer, immense, tremendous, newfound, unparalleled, ...</i>
IMDB	NNSE	<i>neg</i>	217	1.0	<i>debris, trash, garbage, lint, rubbish, sludge, dust, dirt, manure, ...</i>
IMDB	NMF300	<i>pos</i>	100	1.0	<i>imaginative, vivid, lyrical, poetic, realistic, imagery, subtle, ...</i>
IMDB	NMF300	<i>pos</i>	131	0.76	<i>shakira, lauper, mcentire, yearwood, parton, estefan, streisand, ...</i>
BoolQ	SPINE	<i>false</i>	575	1.0	<i>leaked, confidential, libby, fbi, classified, memo, leak, intelligence, ...</i>
BoolQ	SPINE	<i>false</i>	841	0.79	<i>astronaut, soyuz, spacecraft, iss, nasa, astronauts, shuttle, mir, ...</i>
BoolQ	SPOWV	<i>true</i>	758	1.0	<i>cyclone, katrina, hurricane, disaster, ike, flooded, shear, dolly, ...</i>
BoolQ	SPOWV	<i>true</i>	173	0.83	<i>tong, lumpur, myanmar, singaporean, kuala, chung, penang, ...</i>

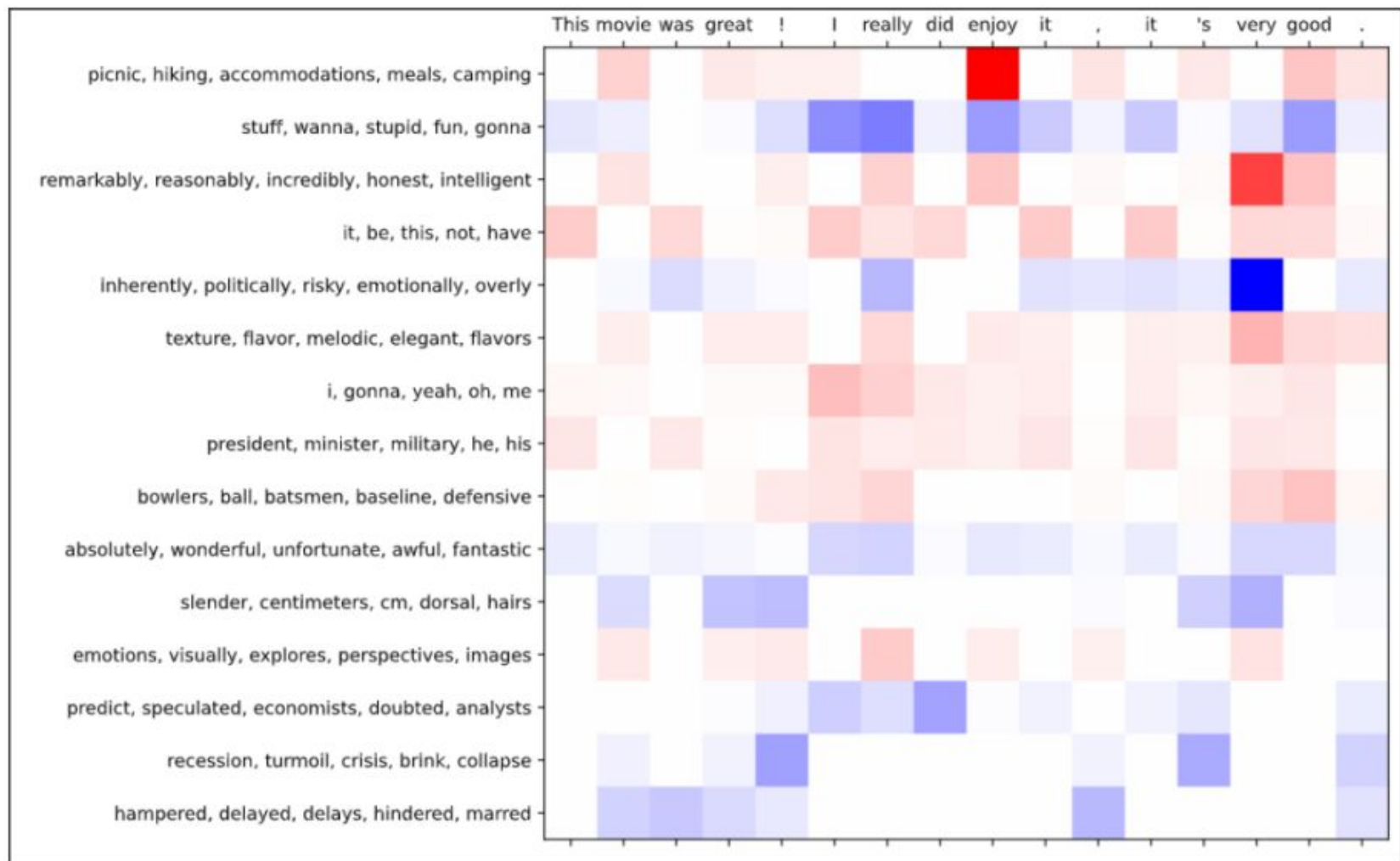


Interprétabilité en TAL - Conclusion Intermédiaire

- Méthode facile à mettre en place, mais pas suffisante seule :
 Σ (Termes Raccourcis) vs. Dimensions parcimonieuses interprétables
- Certaines tâches trop “faciles” : composante lexicale importante
 - Diagnostics de modèles plus complexes : attaques adversariales ?
- Améliorations / Perspectives :
 - Plongements Interprétables “structuraux” vs. “lexicaux” (négation, connecteurs/marqueurs discursifs, etc.)
 - Modèles non-linéaires + Méthodes d’explicabilité

Interprétabilité en TAL - Perspectives

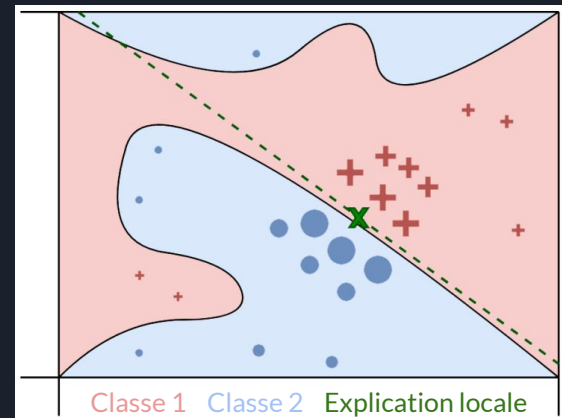
True label	Predicted label	Sentence with saliency ("bettering" "worsening")
Pos	Pos: 1.0	This movie was great ! I really did enjoy it , it 's very good .
Neg	Neg: 0.99951	This movie was awful ! I really did hate it , it 's very bad .
Neg	Pos: 0.99993	This movie was not great ! I really did not enjoy it , it 's not very good .
Pos	Neg: 0.99994	This movie was not awful ! I really did not hate it , it 's not very bad .



Explicabilité en TAL - Contexte :

Méthodes d'Explicabilité

- Explications locales ("contribution" des caractéristiques) :
 - Carte de "saillance"
 - Modèles "substituts" : LIME [\[Ribeiro et al., 2016\]](#)
 - Valeurs de Shapley [\[Shapley, 1953\]](#)
 - Explications "additives" : SHAP [\[Lundberg et Lee, 2017\]](#)



Avantages : Agnostiques aux modèles ; Simples à utiliser/lire

Inconvénients : Difficiles à analyser/interpréter et à mettre en place (en TAL)

- Explications basées sur les exemples :
 - Contrefactuels ("S'il n'y avait pas eu X, il n'y aurait pas eu Y.")
 - Fonctions d'Influence [\[Koh and Liang, 2017\]](#)

Avantages : Plus "naturelles" (causales) ; "Réparation" modèles/jeux de données

Inconvénients : Difficiles à implémenter ; Coûteuses (complexité)



Influence Functions
[\[Koh and Liang, 2017\]](#)

Explicabilité en TAL - Contexte :

Explications en Langage Naturel

PREMISE: *“A girl playing a violin along with a group of people.”*

HYPOTHESIS: *“A girl is washing a load of laundry.”*

GROUND-TRUTH LABEL: *Contradiction*

PREDICTED LABEL: *Contradiction*

FREE-FORM: *“One cannot be playing a violin while washing a load of laundry.”*

e-SNLI

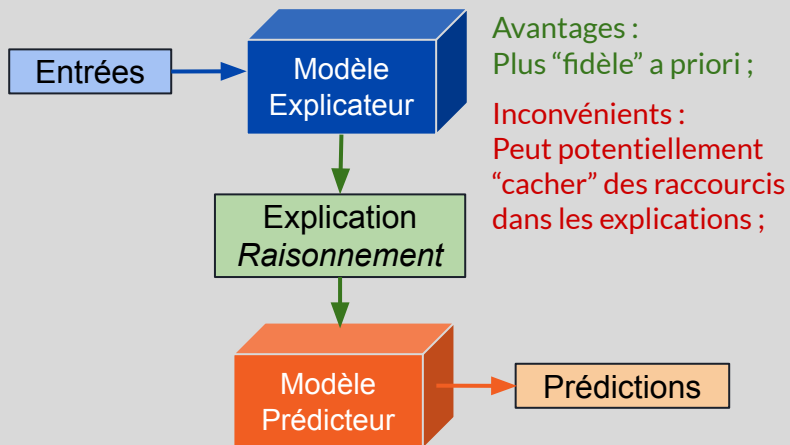
[\[Camburu et al., 2018\]](#)

- Génération d'Explications Naturelles : **Expliquer** et **Prédire** de manière **jointe**
- Comment ?

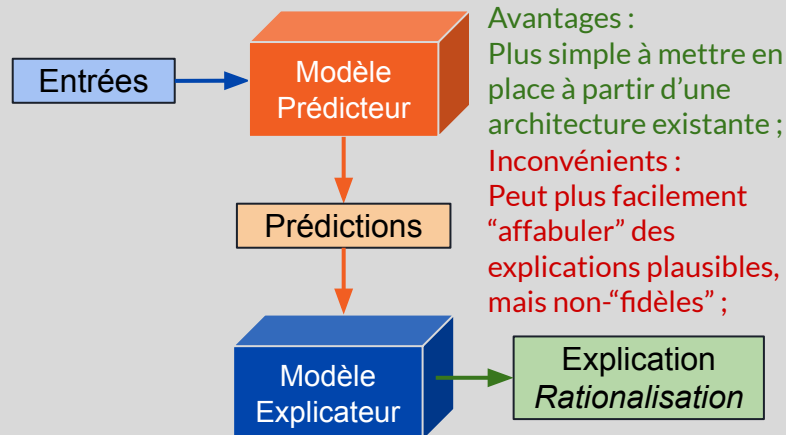
Explicabilité en TAL - Contexte :

Génération Automatique d'Explications

Expliquer-puis-Prédire : "Raisonner"



Prédire-puis-Expliquer : "Rationaliser"

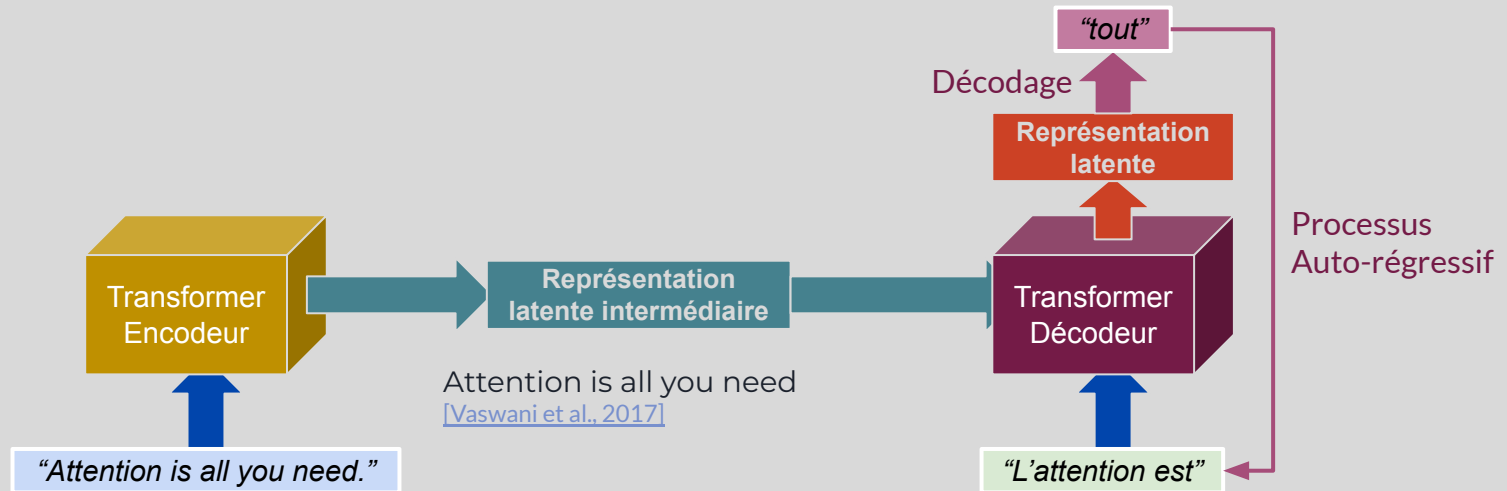


Spectre : Objectifs joints

Explicabilité en TAL - Contexte :

Modèles Transformers - Génération Conditionnée

Génération de Texte Conditionnée



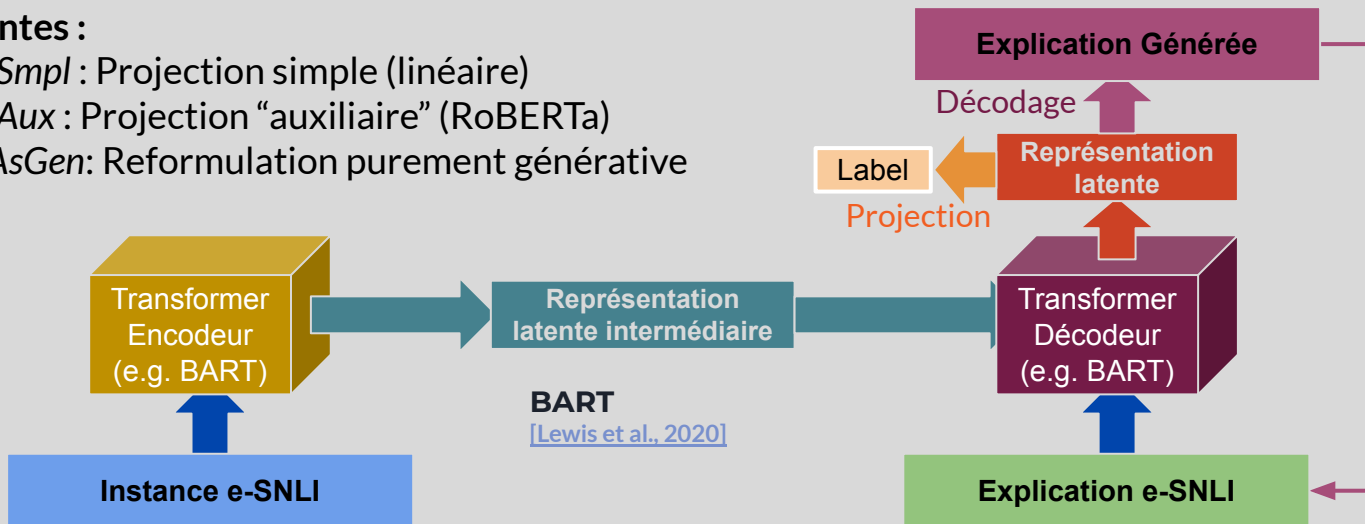
Explicabilité en TAL - Contribution :

Classification et Explication avec Transformer BART

Génération de Texte Conditionnée

Trois variantes :

- *JointSmpl* : Projection simple (linéaire)
- *JointAux* : Projection "auxiliaire" (RoBERTa)
- *ExplAsGen*: Reformulation purement générative





Explicabilité en TAL - Contributions :

Évaluation - Tâches en Aval

E-SNLI					
Model	Accuracy	BLEU	ROUGE-1-F1	ROUGE-2-F1	ROUGE-L
JointSmpl	83.03	17.8	42.87	22.15	38.51
JointAux	70.27	18.66	43.82	23.21	39.51
ExplAsGen	91.03	28.05	58.67	38.46	54.9
Camburu et al.	81.71	27.58	-	-	

Explicabilité en TAL - Contributions :

Évaluation - Qualité des Explications

Premise	Hypothesis	Gold Label	Output A	Output B
Two men prepare a fish at a dock.	Two men are cleaning their fish	Entailment	Neutral , because preparing a fish is not cleaning their fish.	Neutral , because preparing a fish does not imply cleaning their fish.
A yellow dog is running in a field near a mountain.	A yellow dog is going to the vet.	Contradiction	Contradiction , because running and going are not the same.	Neutral , because a yellow dog is running in a field near a mountain does not indicate that it is going to the vet.

- Comparaisons deux-à-deux entre **Sortie A** et **Sortie B**, sur trois critères :
 - La “fluidité” (**Fluency**) : Est-ce que l’explication est lisible et écrite correctement ?
 - La pertinence et la “couverture” (**Relevance and Coverage**) : Est-ce que tout les éléments pertinents de l’instance sont mentionnés dans l’explication ?
 - L’utilité (**Utility**) : Est-ce que l’explication produite est utile à la compréhension de la décision du modèle ?

Explicabilité en TAL - Contributions :

Évaluation - Qualité des Explications

Model	Fluency	Relevance & Coverage	Utility
JointSmpl	12	22	26
JointAux	14	26	16
ExplAsGen	24	13	21
<i>Indecision</i>	40	29	27
Fleiss' Kappa	0.44	0.59	0.49

Model	Fluency	R&C	Utility
JointSmpl	26	43	49
Camburu et al.	23	10	12
<i>Indecision</i>	41	37	29
Fleiss' Kappa	0.17	0.15	0.47

Explicabilité en TAL - Contributions :

Évaluation - Qualité des Explications

Premise	Four children are playing in some water.
Hypothesis	The children are wet.
Gold label	Entailment
JOINTSMPL	Entailment, because children are playing in some water is same as children are wet.
JOINTAUX	<u>Neutral</u> , because playing in some water does not imply being wet.
EXPLASGEN	Entailment, because the children are playing in water so they must be wet.
Camburu et al.	Entailment, because children playing in water are wet.
Gold Explanation 1	Entailment, because playing in water means you are wet.
Gold Explanation 2	Entailment, because the children became wet as they are playing in water.
Gold Explanation 3	Entailment, because four children are children, and playing in water implies wet.



Explicabilité en TAL - Conclusion Intermédiaire

- Méthodes d'explicabilité existantes difficiles à mettre en place en TAL (interprétabilité)
- Génération d'Explications Naturelles :
 - Approche pertinente compte tenu de l'évolution des modèles génératifs en TAL :
 - GPT-3 [\[Brown et al., 2020\]](#) / T5 [\[Raffel et al., 2020\]](#) : *One/Few-Shot Learning, Prompt Engineering*
 - Limites et problèmes :
 - Difficile de garantir la fidélité des explications : “Épreuves” et évaluations adversarielles ([\[Camburu et al., 2020\]](#))
 - Évaluation de la qualité des explications : protocoles plus “formels” pour les évaluations humaines ; meilleurs métriques automatiques

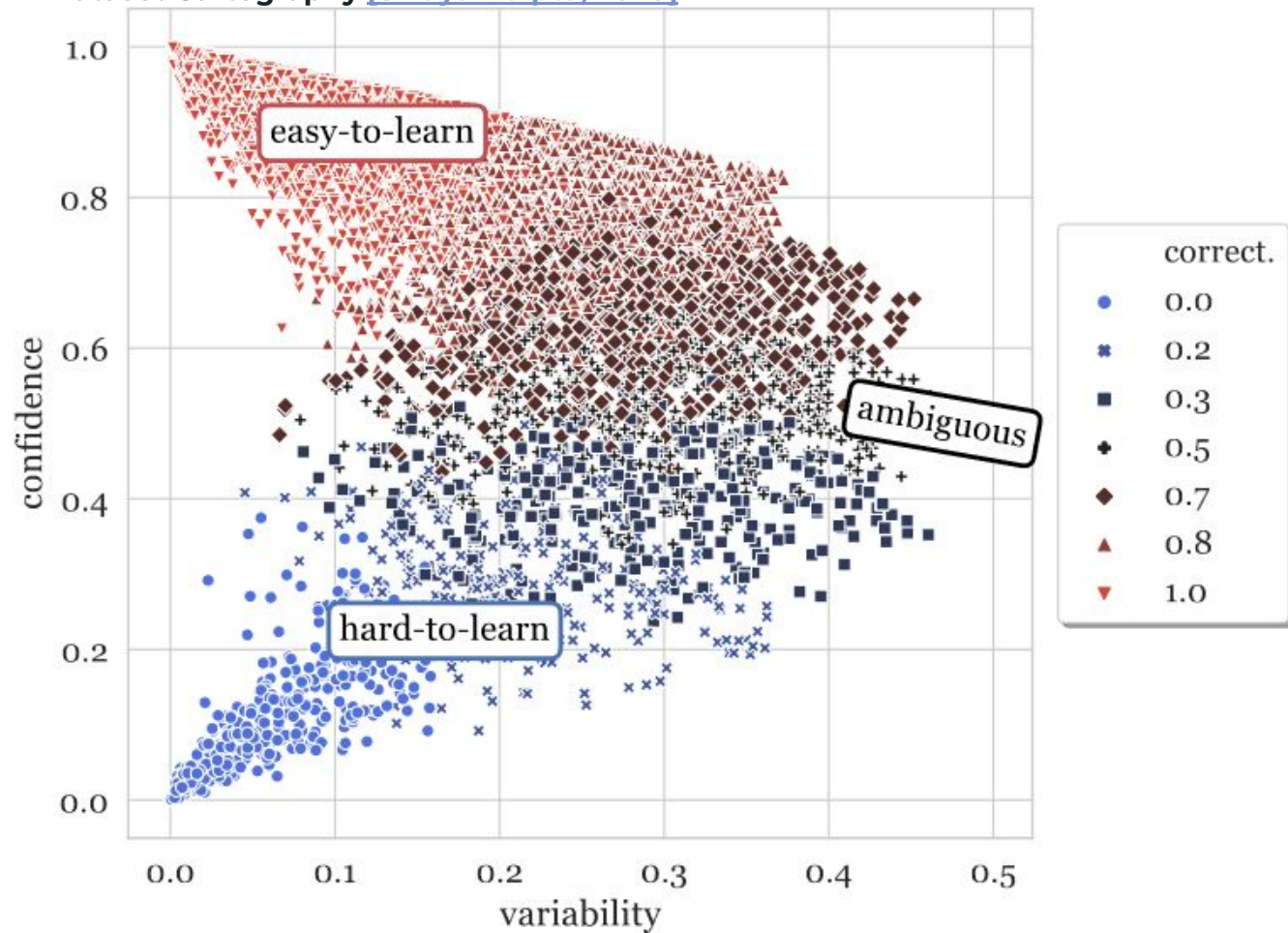
- 



Merci de votre
attention

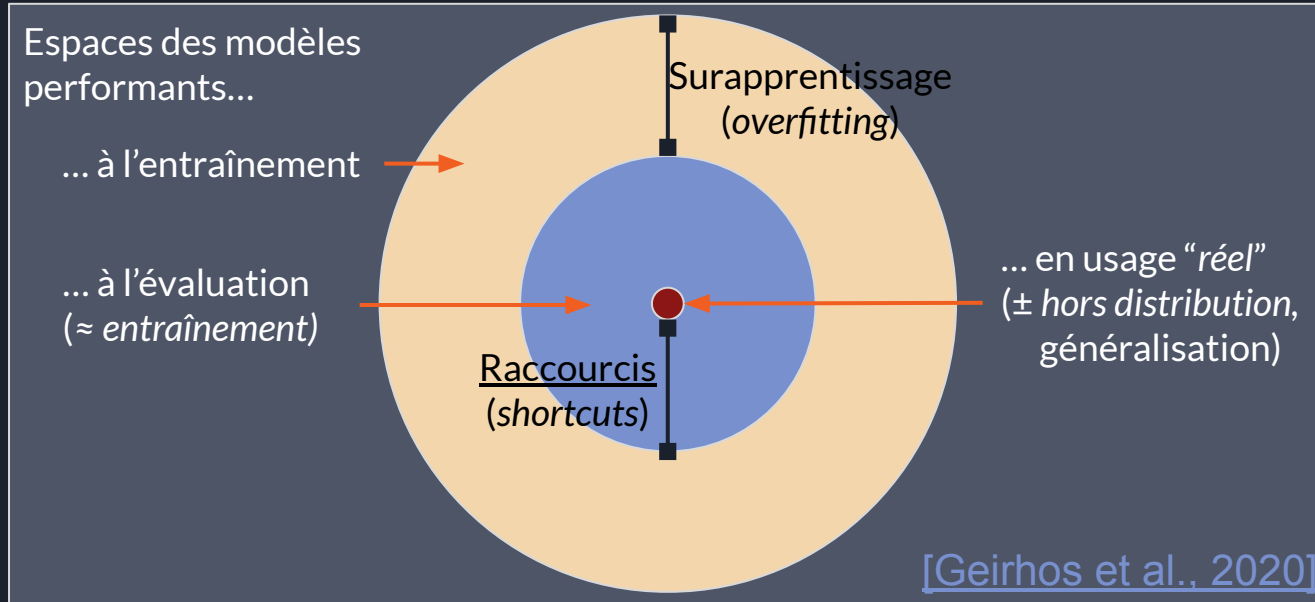
Des Questions ?

Dataset Cartography [Swayamdipta, 2020]



Contexte : Apprentissage par “raccourcis”

Principe du Moindre Effort $\Rightarrow \text{Coût}(\sum \text{Raccourcis}) < \text{Coût}(\text{Généralisation des données})$





Contexte : Problèmes de construction/collection

- Besoins de grandes quantités de données $\Rightarrow$ Myriadisation (*crowdsourcing*)
- “Facilités” d’annotation :

Premise	People are waiting in line in a snowstorm.	
Entailment	People are in a <u>snowstorm</u> .	
Contradiction (Good)	People are waiting in a line during a <u>scorching drought</u> .	
Contradiction (Shortcut)	People are not waiting in line in a snowstorm.	[are $\rightarrow$ are not]
Contradiction (Shortcut)	Nobody is waiting in line in a snowstorm.	[People $\rightarrow$ Nobody]

Interprétabilité en TAL - Contexte :

Modèles existants

NNSE

[Murphy et al., 2012]

Contraintes A
Posteriori

SPOWV

[Faruqui et al., 2015]

SPINE

[Subramanian et al., 2018]

Factorisation
Non-Négative

NMF300

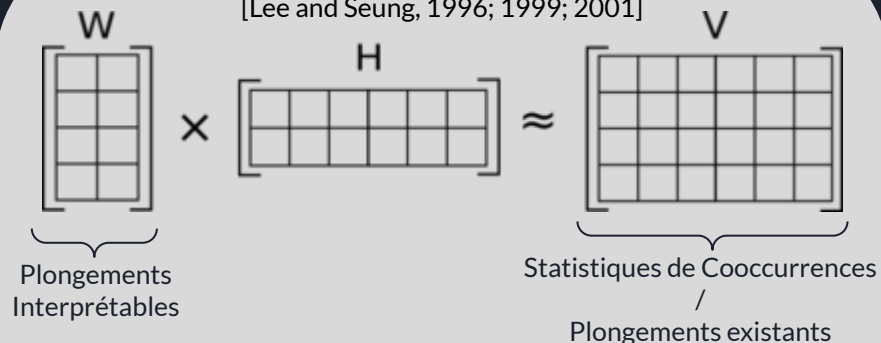
[Bourgeade et al., 2021]

Word2Sense

[Panigrahi et al., 2019]

Factorisation Non-Négative

[Lee and Seung, 1996; 1999; 2001]

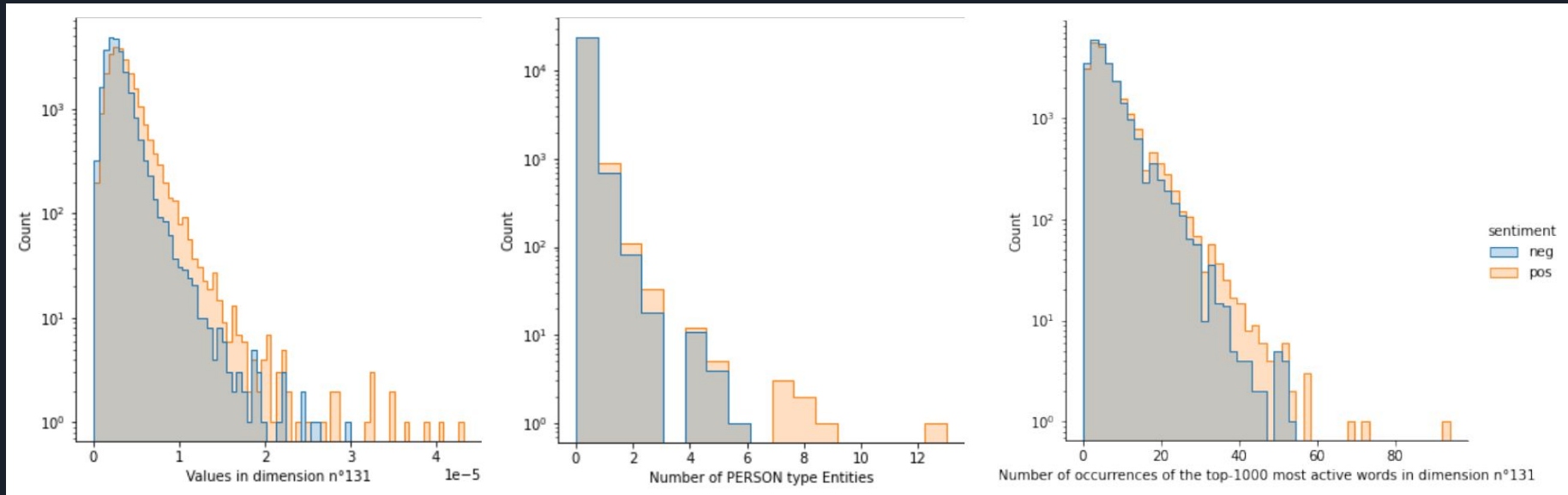


sous contraintes : W et/ou $H \geq 0$ + contrainte de parcimonie (surcomplétude; non-unicité)

Coût(Plongements Denses)
 $\approx$ **Coût(Plongements Non-Négatifs Parcimonieux)**

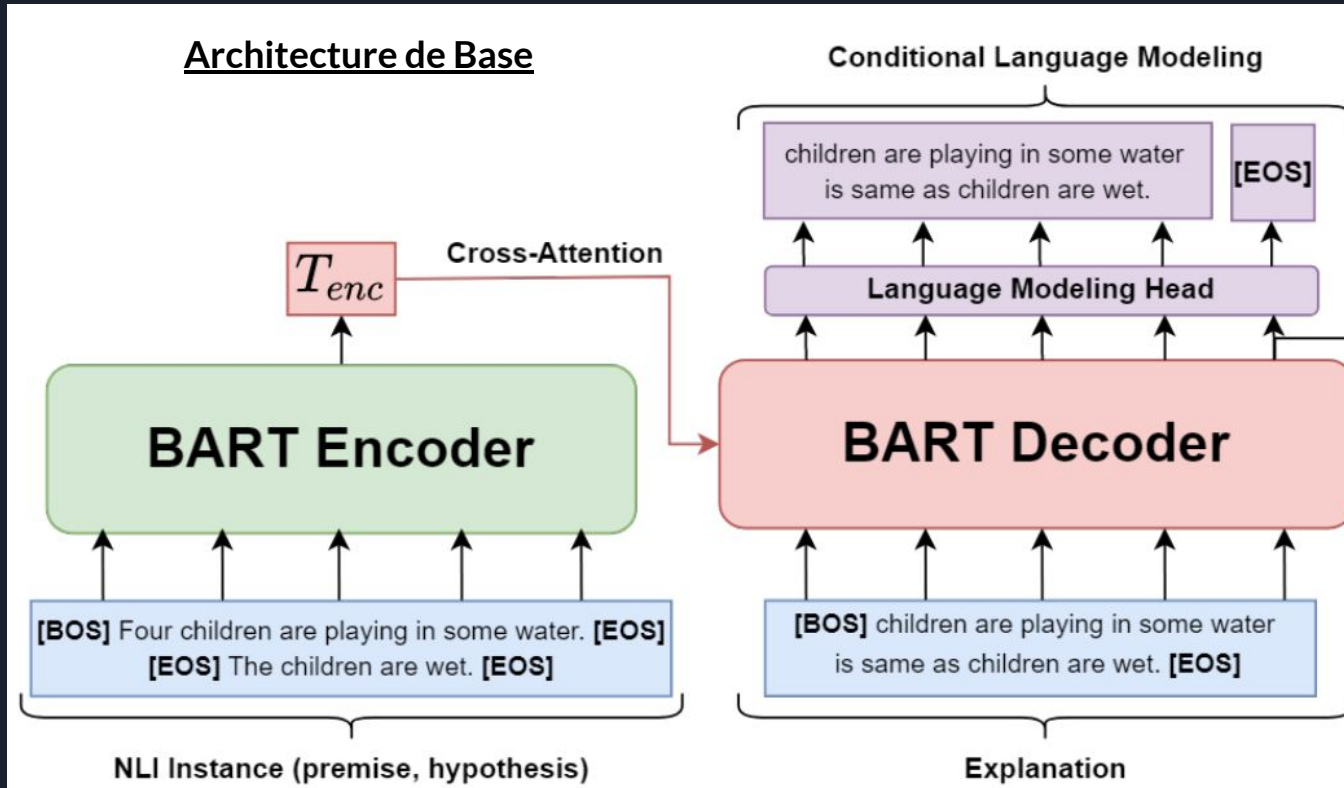
Interprétabilité en TAL - Contributions :

Évaluation - Diagnostics de Corpus



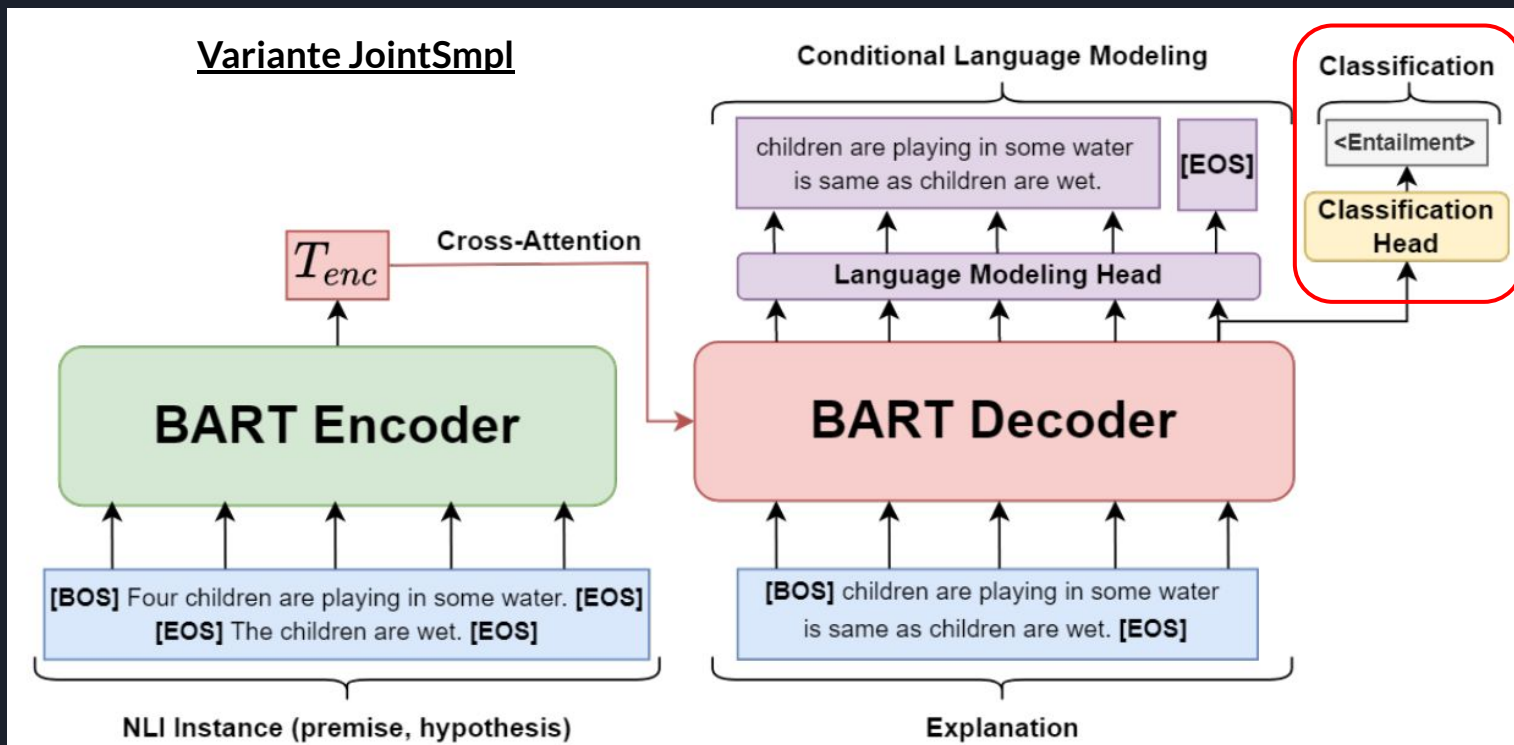
Explicabilité en TAL - Contributions :

Modèles Classifieur-Explicateur



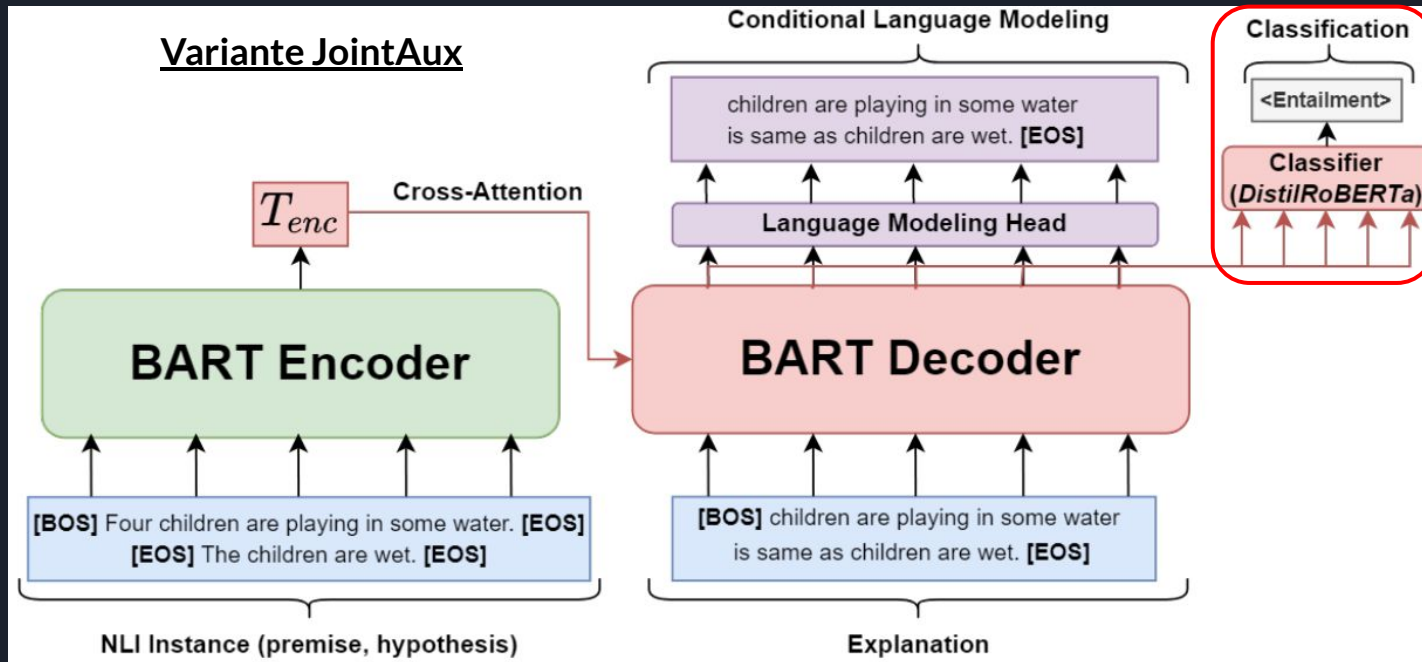
Explicabilité en TAL - Contributions :

Modèles Classifieur-Explicateur



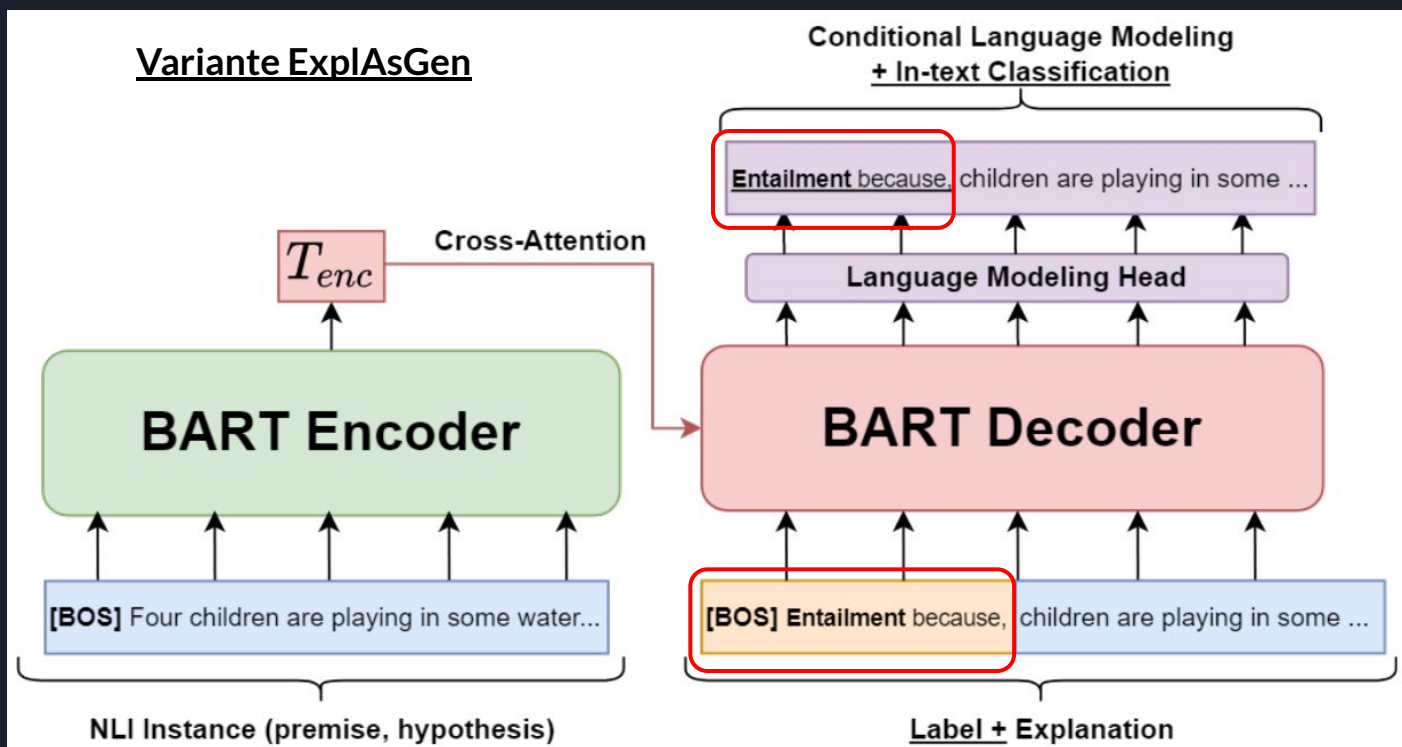
Explicabilité en TAL - Contributions :

Modèles Classifieur-Explicateur



Explicabilité en TAL - Contributions :

Modèles Classifieur-Explicateur



Explicabilité en TAL - Contributions :

Évaluation - Qualité des Explications

Premise	Two dogs are playing catch in a field.
Hypothesis	The two dogs are resting in the field.
Gold label	Contradiction
JOINTSMPL	Contradiction, because playing and resting are not the same.
JOINTAUX	Contradiction, because the dogs cannot be playing catch and resting at the same time.
EXPLASGEN	Contradiction, because the dogs cannot be playing catch and resting simultaneously.
Camburu et al.	Contradiction, because the dogs are either playing catch or resting.
<i>Gold Explanation 1</i>	Contradiction, because if the two dogs are playing, they can not be resting.
<i>Gold Explanation 2</i>	Contradiction, because one cannot be playing with resting simultaneously.
<i>Gold Explanation 3</i>	Contradiction, because “ playing ” and “ resting ” contradict one another in that playing shows movement while resting shows non-movement.



Explicabilité en TAL - Contributions :

Évaluation - Qualité des Explications

Premise	Four children are playing in a pool in swimsuits .
Hypothesis	The children are wet.
Gold label	Entailment
JOINTSMPL	Entailment, because the children are playing in a pool so they must be wet.
JOINTAUX	<u>Neutral</u> , because just because children are playing in a pool in swimsuits does not mean they are wet.
EXPLASGEN	Entailment, because children playing in a pool are wet.