Explaining semantic representations in natural language processing

Philippe Muller, Tom Bourgeade

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE TOULOUSE INSTITUTE

IRIT
Institut de Recherche en Informatique de Toulouse
CNRS · INP · UT3 · UT1 · UT2J

MELODI

Context

NLP today:

- considerable recent progress on a lot of tasks
- mostly based on learned “semantic” representations
- relies on pretrained models
  + task specific fine-tuning

Problems:

- robustness
- biased productions
- exploiting artifacts

Overview

Problems:

- robustness
- biased productions
- exploiting artifacts

Potential solutions: XAI methods

- interpreting a model's behaviour
- explaining model's decisions

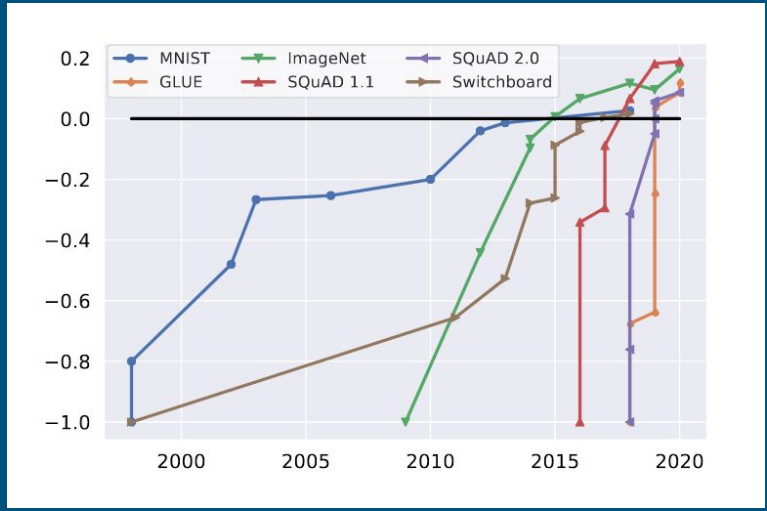
Specificities of NLP models
- nature of the input
- modes of explanation

Example studies
- interpretable representation
- self-explaining model

Modern NLP progress

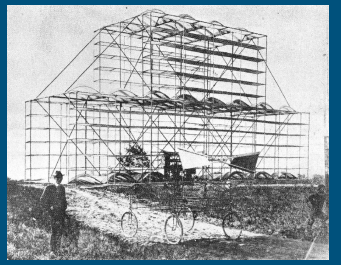
- quick saturation of benchmarks
- complexity of models

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



Kelia et al., 2021, Dynabench: Rethinking Benchmarking in NLP

But: Models are not robust



VQA

Original	What color is the flower ?
Reduced	flower ?
Answer	yellow
Confidence	0.827 → 0.819

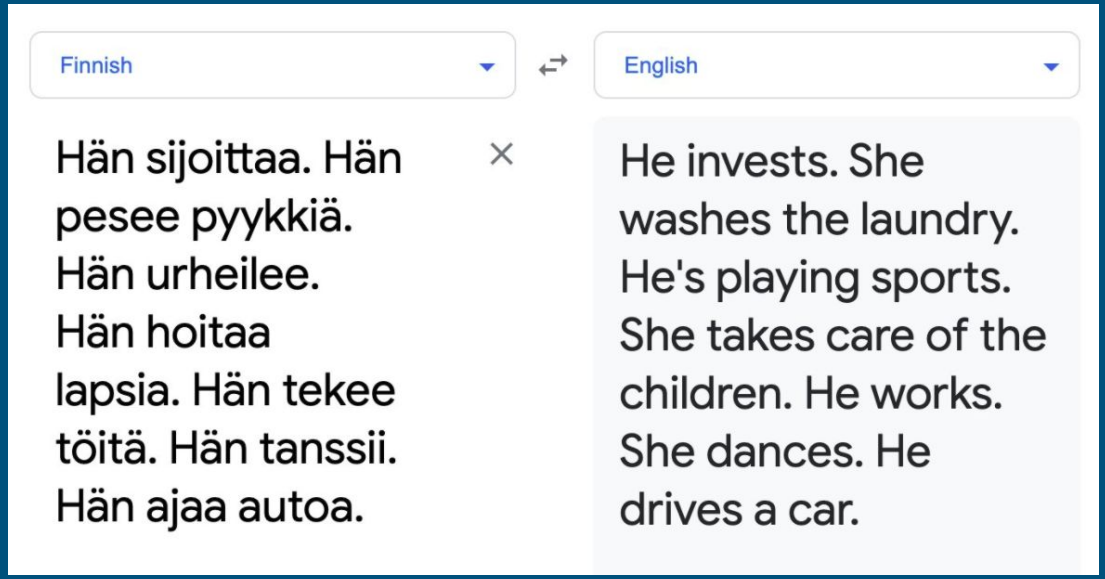
Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, Jordan L. Boyd-Graber: Pathologies of Neural Models Make Interpretation Difficult. EMNLP 2018

Original
Perfect performance by the actor → **Positive (99%)**

Adversarial
Spotless performance by the actor → **Negative (100%)**

Morris et al., 2020,
TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. EMNLP 2020

But: Models are biased



(Viral example by Vuokko Aro)

see also : Sheng, Change, Natarajani & Peng “The Woman Worked as a Babysitter: On Biases in Language Generation”, EMNLP 2019

But: Models pick up on artifacts

The way datasets are collected is prone to artificial peculiarities

Example: SNLI (Stanford Natural Language Inference) [Bowman et al., 2015]: can the hypothesis be inferred from the premise ?

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.

But: Models pick up on artifacts

The way datasets are collected is prone to artificial peculiarities.

Example: SNLI (Stanford Natural Language Inference) [Bowman et al., 2015]: can the hypothesis be inferred from the premise ?

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.

Hypothesis-only model ⇒ ~ 67 % accuracy [Poliak et al., 2018]

Models pick up on artifacts

"Universal triggers": add a string to all inputs, that switches the decision.

Ex on SNLI: nobody, never, switch almost 100% of instances to "contradiction"

> Premise: a boy and girl are playing.
> Hypothesis: nobody two people are playing outside.

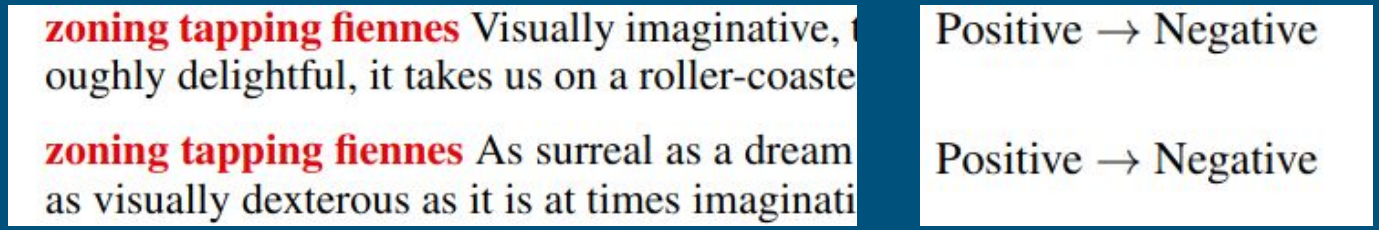
→artifact in contradiction examples, easy !

Wallace et al., 2019: Universal Adversarial Triggers for Attacking and Analyzing NLP

Models pick up on artifacts

"Universal triggers": add a string to all inputs, that switches the decision.

Example on sentiment analysis

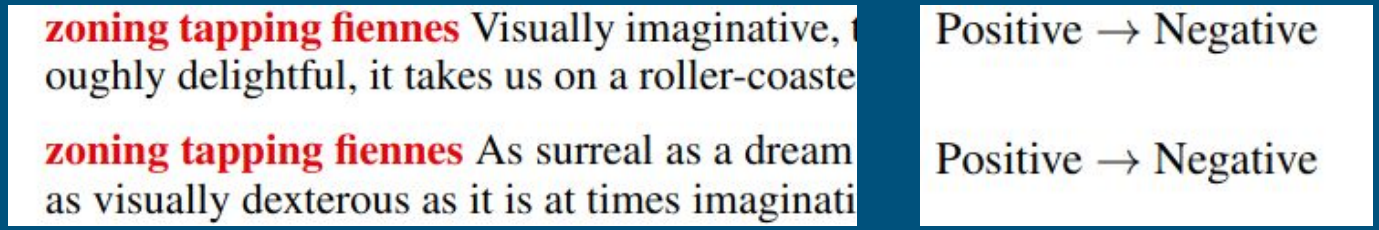


Wallace et al., 2019: Universal Adversarial Triggers for Attacking and Analyzing NLP

Models pick up on artifacts

"Universal triggers": add a string to all inputs, that switches the decision.

Example on sentiment analysis ???



Wallace et al., 2019: Universal Adversarial Triggers for Attacking and Analyzing NLP

Why is natural language specific ?

Language “operates” on a discrete, sparse space (words):

I enjoyed this movie.

I didn’t enjoy this movie.

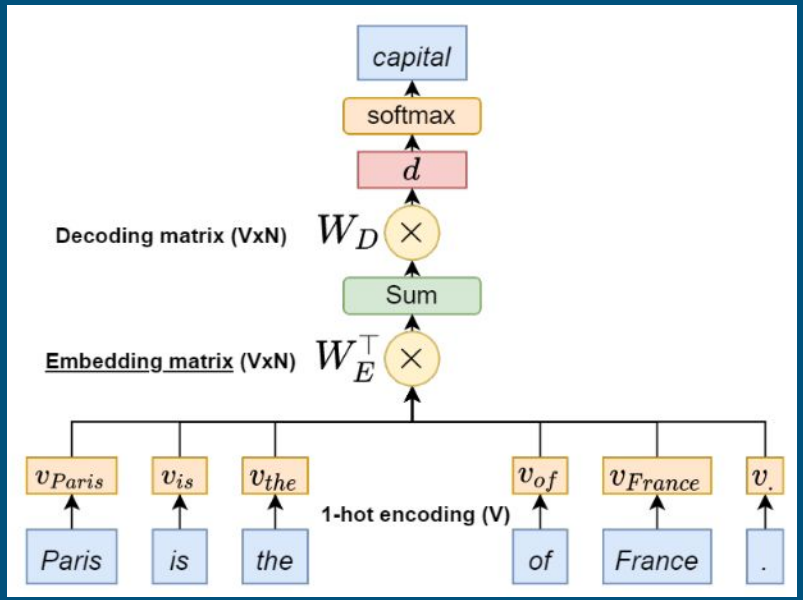
I didn’t like this movie.

This film was so bad.

A small change can make a big difference… or not.

Different “surface” forms can have very similar senses

Vectorial representation for words



Example: Word2vec (CBOW version) — 2013

- learns to predict a word given a small context
- makes the network learn vectorial representations for words (“embeddings”)
- words occurring in similar contexts will have similar representations

approximates a type of “semantic” similarity

Why is natural language specific ?

Language has underlying structures, but is expressed in a sequence

The movie I went to see with a friend was so horrible.

- i went to see a movie with a friend
- the movie was horrible

Why is natural language specific ?

Language has underlying structures, but is expressed in a sequence

The movie I went to see with a friend was so horrible.

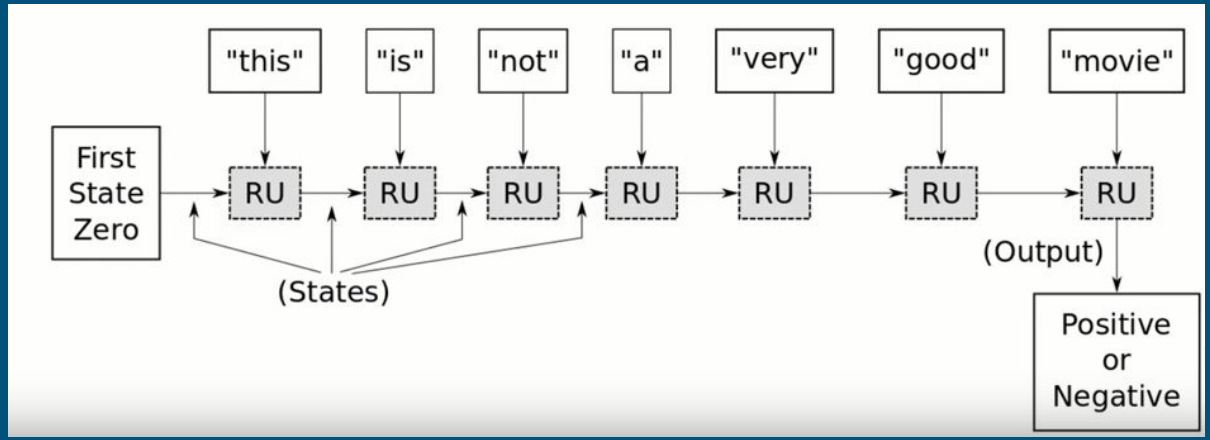
- i went to see a movie with a friend
- the movie was horrible

Sentence level representations ?

How to compose word representations into a sentence ?

Typical solution 5 years ago: a recurrent neural network

- Task specific
- Brutal aggregation (cf Ray Mooney's famous quote)
- Long distance dependencies are lost



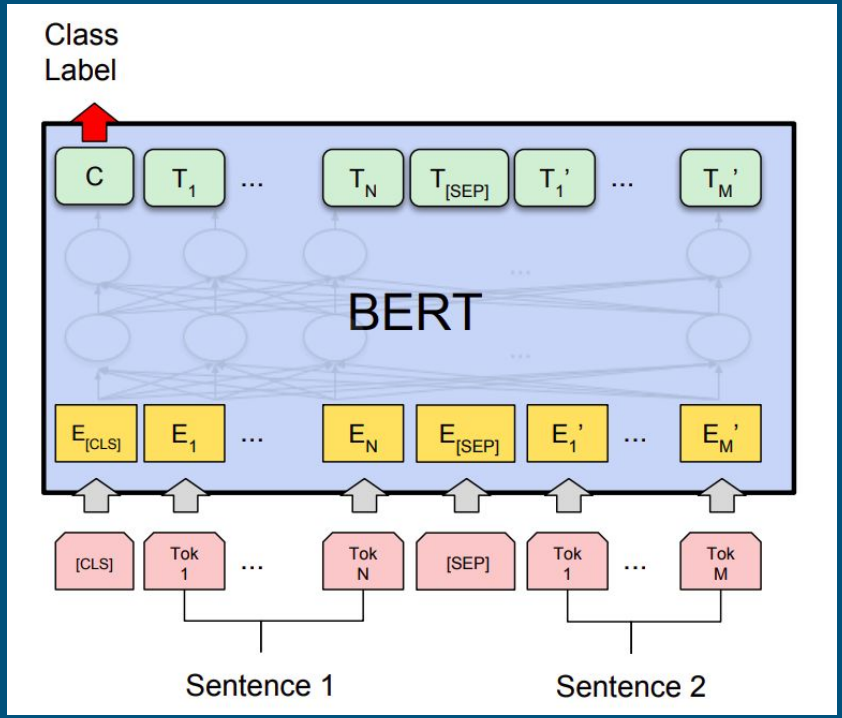
Sentence level representations ?

Typical solution today:

- contextual embeddings, pretrained (BERT & Co)

Contextual embeddings:

- training ~similar to word2vec, but with whole sentential context
- deep interdependent representations
- the same architecture is used for downstream tasks, only fine-tuned



Challenges for explainable methods on NL

- Word embeddings : models operates on the embedded space → harder to link to original inputs (compared to e.g. images)
- Sequential models: compositions are hard to trace
- Contextual embeddings: deep distributed representations ; added complexity

Approaches for explainable models

- design transparent models / built-in interpretability

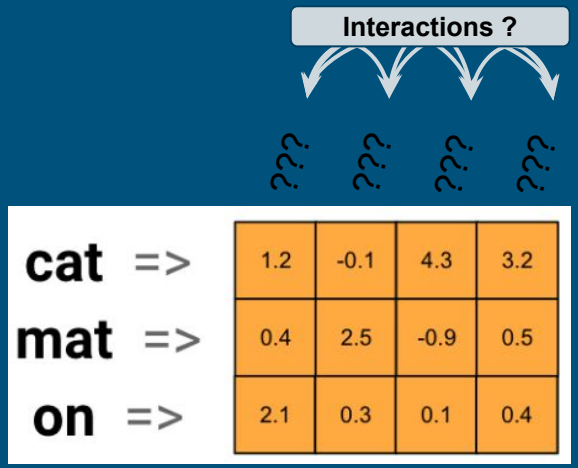
- black-box :
  - model inspection / probing
  - model outcome explanation
  - global model explanation

Approaches for explainable models

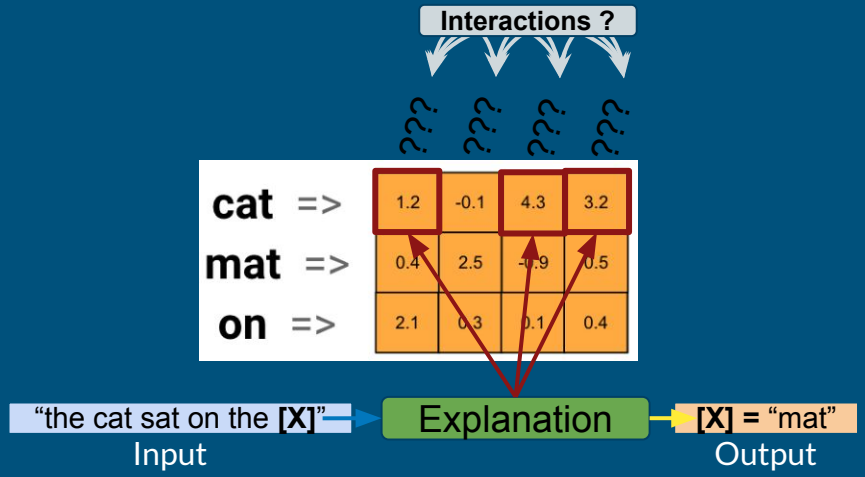
- design transparent models / built-in interpretability
  Case study : design interpretable word embeddings

- black-box :
  - model inspection / probing
  - model outcome explanation
  - global model explanation

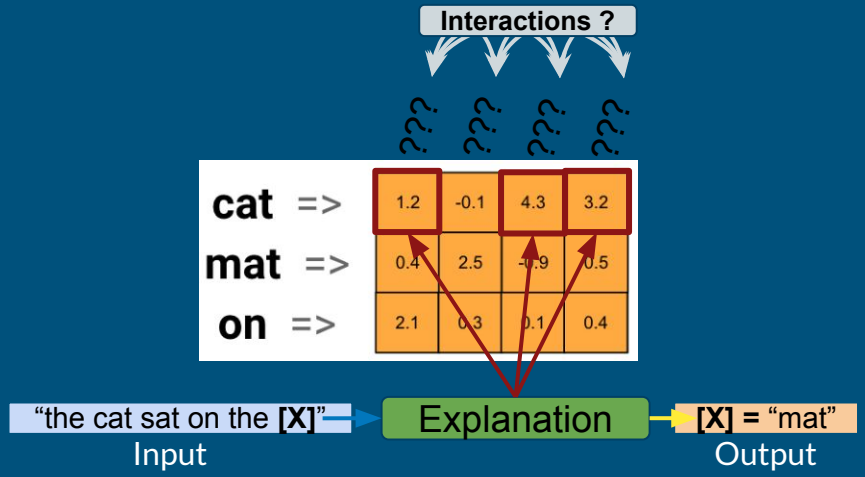
What’s in an embedding ?



Interpreting the dimensions ?



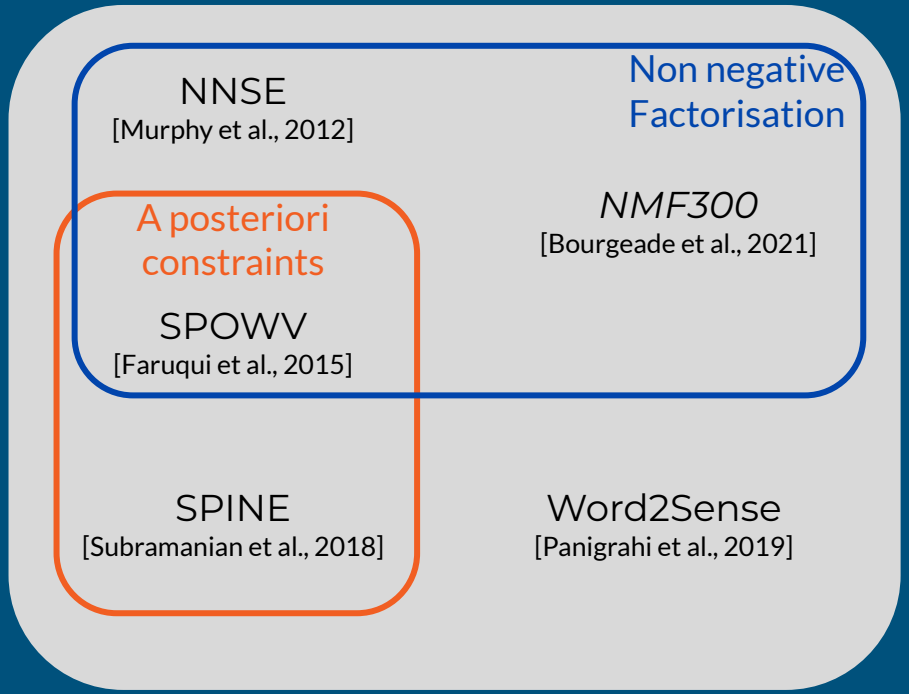
Interpreting the dimensions ?



Objective: sparsity+non negativity

Dimensions Interprétables	cat	dog	bat	mat
cat, dog, bird, bat …	0.6	0.6	0.5	0
feline, catlike, stealthy, …	0.2	0	0.2	0
canine, hound, puppy, …	0	0.2	0.3	0
pet, domesticated, tamed, …	0.2	0.2	0	0
fabric, cloth, leather, …	0	0	0	1.0

Interpreting the dimensions ?



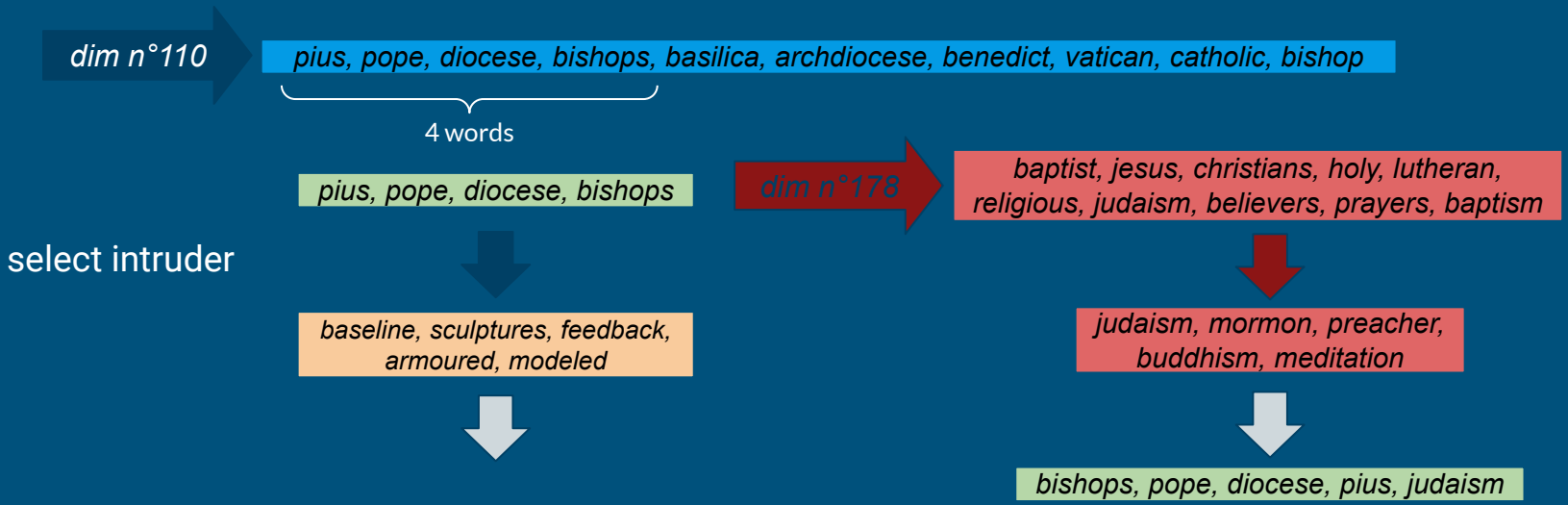
	NMF300
science	neurology, ophthalmology, oncology, radiology, microbiolo courses, curriculum, undergraduate, students, vocationa harvard, phd, doctorate, caltech, yale, swarthm

	word2vec
science	insult, rivalries, reactors, mw, armistic editing, airplay, cds, professionally, songwri ss, rbi, viii, 2d, xiii, shortst

Consider the most active dimensions for a word
Look at word most active in that dimension

“Measuring” interpretability : word intrusion

[Chang et al., 2009]



“Measuring” interpretability : word intrusion

[Chang et al., 2009]

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

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

Approaches for explainable models

- design transparent models / built-in interpretability
  Case study : design interpretable word embeddings

- black-box :
  - model inspection / probing
  - model outcome explanation
  - global model explanation

Approaches for explainable models

- design transparent models / built-in interpretability
  Case study : design interpretable word embeddings

- black-box :
  - model inspection / probing
  - model outcome explanation
  - global model explanation

model specific

Approaches for explainable models

- design transparent models / built-in interpretability
  Case study : design interpretable word embeddings

- black-box :
  - model inspection / probing — model specific
  - model outcome explanation
  - global model explanation — very hard

Model outcome explanation

The most popular approach … many different methods:

- input attribution : which part of the input influenced the decision (again many variants)

- local approximations (LIME, Anchor,...): disturb an instance to generate a neighborhood, apply an interpretable model on it (linear, explicit rules, …)

Model outcome explanation

The most popular approach … many different methods:

- input attribution : which part of the input influenced the decision (again many variants)

  For NL, which relevant part ? words, phrases, … ? Cf robustness problems

- local approximations (LIME, Anchor,...): disturb an instance to generate a neighborhood, apply an interpretable model on it (linear, explicit rules, …)

Model outcome explanation

The most popular approach … many different methods:

- input attribution : which part of the input influenced the decision (again many variants)

  For NL, which relevant part ? words, phrases, … ? Cf robustness problems

- local approximations (LIME, Anchor,...): disturb an instance to generate a neighborhood, apply an interpretable model on it (linear, explicit rules, …)
  For NL, generating a “neighbour” is far from obvious

Model outcome explanation

The most popular approach … many different methods:

- input attribution : which part of the input influenced the decision
  (again many variants)

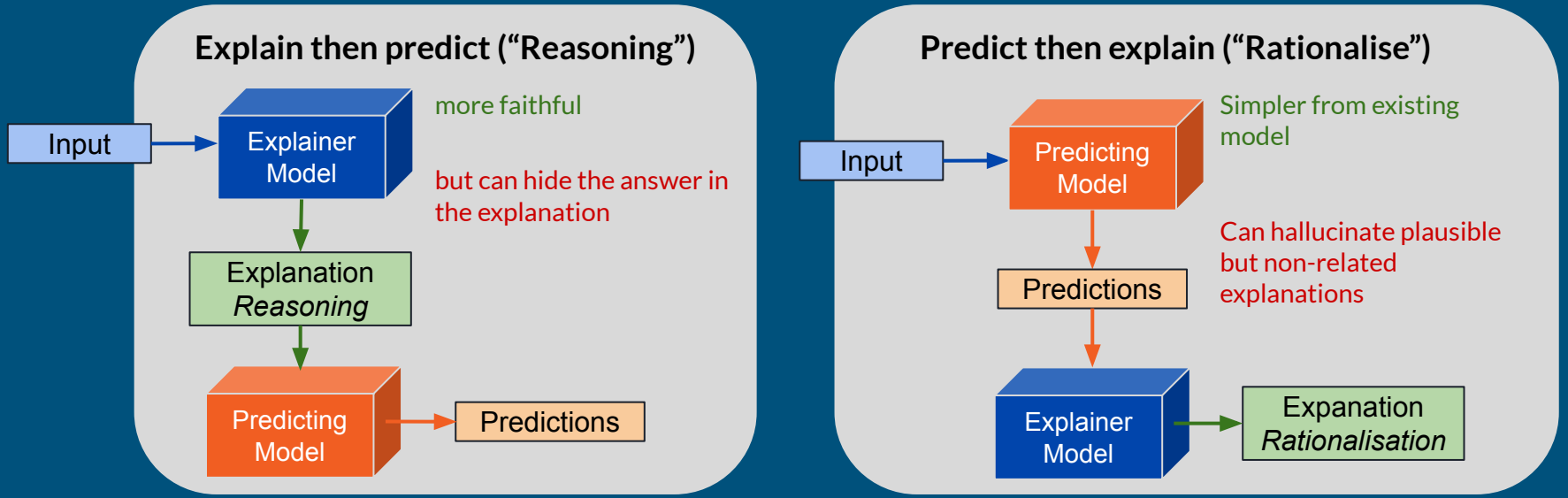
  For NL, which relevant part ? words, phrases, … ? Cf robustness problems

  How about a whole sentence ?

- local approximations (LIME, Anchor,...): disturb an instance to generate a neighborhood, apply an interpretable model on it (linear, explicit rules, …)
  For NL, generating a “neighbour” is far from obvious

NLP Self-explaining models

Jointly classify and generate a natural language explanation that uses the input



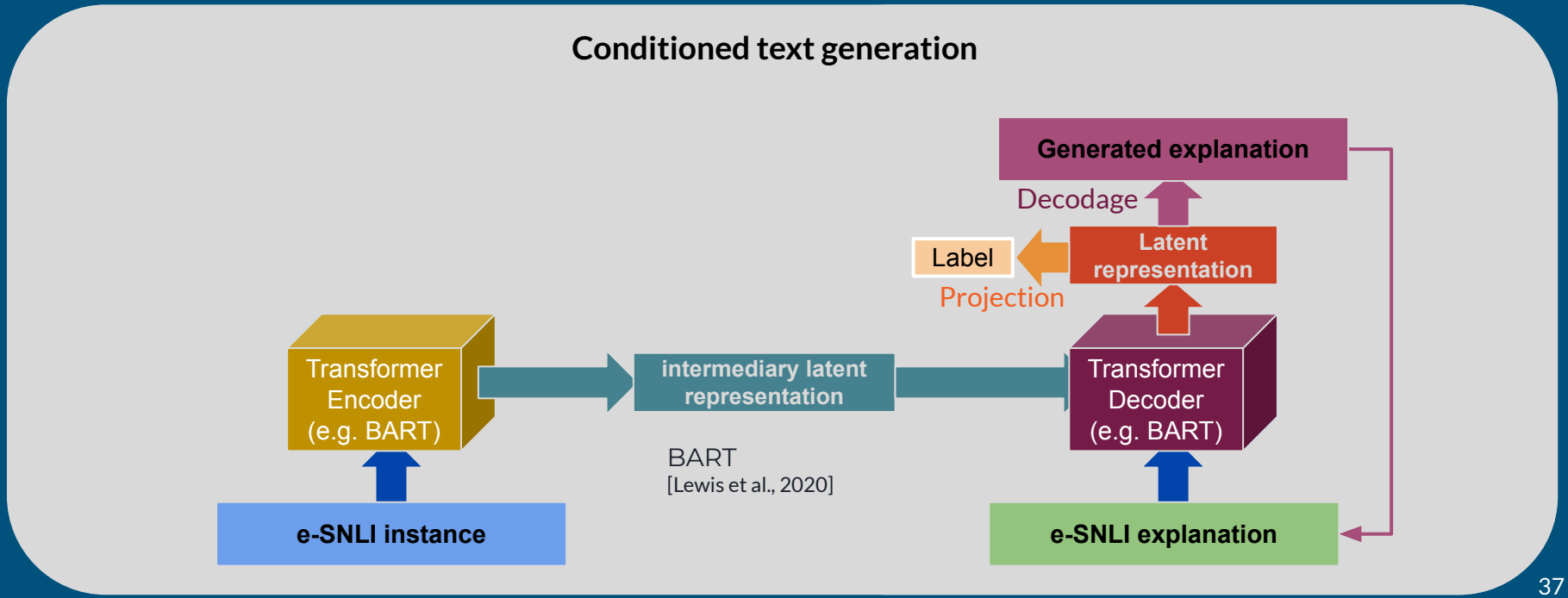
Experiments: Natural Language Inference

Using the SNLI corpus, and additional explanation annotation :
e-SNLI [Camburu et al., 2018]

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]

Architecture: sequence-to-sequence



Evaluation: automated metrics

	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	-	-	

NB: best current accuracy for classif model ~ 92% (Pilault et al., 2021)

Evaluation: manual comparison

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.

- Fluency
- Relevance & Coverage: explanation includes all and only relevant parts of the inputs
- Utility: overall quality of the explanation

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

Example output

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	Neutral, 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.

Conclusion

- NLP has made progress due to powerful intermediate pretrained representations, but it’s unclear what these capture linguistically
- Applying explainable methods is challenging on NLP problems
- NL explanations need a lot of additional data
- Depends on the objective/use case/end user ?

  debugging, understanding, persuasion, control, …