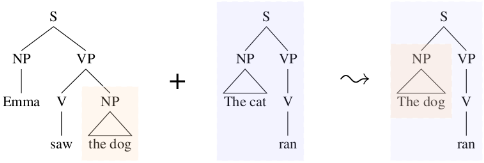
COMPO: Structural Biases for Compositional Generalisation

Maximin Coavoux

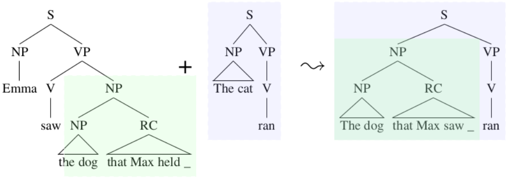
maximin.coavoux@univ-grenoble-alpes.fr

Challenge datasets for compositional generalization

- COGS (Kim et al 2020) $\longrightarrow$ SLOG (Li et al 2023): artificial semantic parsing datasets constructed to assess the **compositional generalization** abilities of NLP models
- Systematic shifts / gaps between train and test



(a) Lexical generalization: object → subject (COGS)



(b) Structural generalization: RC object→RC subject (SLOG)

Emma saw the dog

- Clasical format:
  `*dog(x3); see.agent(x1, Emma) AND see.theme(x1, x3)`
- Variable free format:
  `see(agent=Emma, theme=*dog)`

Challenge datasets for compositional generalization

- **COGS**:
  - Lexical generalization 80 % of the corpus
  - Structural generalization : 20 %
  - seq2seq models at around 80 % accuracy (100 on lexical, ~0 on stuctural)
- **SLOG**: focus on structural generalization
  - larger coverage of syntactic structures
  - center embedding, wh-questions, relative clauses
  - larger coverage of generalization types :
    - ★ generalization to shorter or deeper recursion
    - ★ distribution of semantic roles

Challenge datasets for compositional generalization

- **COGS**:
  - Lexical generalization 80 % of the corpus
  - Structural generalization : 20 %
  - seq2seq models at around 80 % accuracy (100 on lexical, ~0 on stuctural)
- **SLOG**: focus on structural generalization
  - larger coverage of syntactic structures
  - center embedding, wh-questions, relative clauses
  - larger coverage of generalization types :
    - generalization to shorter or deeper recursion
    - distribution of semantic roles
- A limitation of COGS / SLOG from COMPO's perspective:
  - The type of generalization expected in COGS model some form of language **competence** (extremely deep recursion) that might not be appropriate if we take into account speakers' memory limitations
  - `Gabriel enlarged a donut that the teacher that a spokesman that the lawyer that a bird that a girl that a boy that a squirrel that Emma ate helped rolled packed drew floated tolerated cleaned .`

COGS and syntax: Yao and Keller (2022)

- Structural generalization is hard for seq2seq models
  - BART / T5 fail on structural generalisation
  - even when given gold syntactic trees as input
- What happens if you simplify the task ?
  1. replace Logical Form by gold syntactic tree
     - ★ `( S ( NP ( Det The ) ( N baby ) [......] ( VP ( V screamed ) ) )`
  2. replace Logical Form by sequence of POS
     - ★ `Det N P Det N P Det N V`
  - BART / T5 still fail (a bit less miserably)
- In contrast, Berkeley Neural parser is 84-99 % (depending of types of syntactic generalization)

COGS and syntax: Yao and Keller (2022)

	Model Class	Model	Struct: Obj to Subj PP	Struct: CP recursion	Struct: PP recursion	Lex: all 18 other types	Overall
semantics	seq2seq	BART	0	0	12	91	79
		BART+syn	0	5	8	93	80
		T5	0	0	9	97	83
		Kim and Linzen 2020	0	0	0	73	63
		Akyürek and Andreas 2021	0	0	1	96	82
		Zheng and Lapata 2022	0	12	39	99	89
		Conklin et al. 2021	0	0	0	88	75
		Csordás et al. 2021	0	0	0	95	81
		Qiu et al. 2021 *	100	100	100	100	100
	structure-aware	Liu et al. 2021	93	100	99	99	99
		Weißenhorn et al. 2022	78	100	99	100	98
syntax	seq2seq	BART	0	9	22	99	87
		T5	5	7	9	99	86
	structure-aware	Neural Berkeley Parser	84	95	98	100	99
POS tags	seq2seq	BART	0	6	19	98	85
		T5	0	4	4	98	85
	structure-aware	most frequent POS	92	98	100	92	93

COGS and syntax : Lessons Learned

Strong mismatch between the task and the tool

If the only tool you have is ~~a hammer~~ GPT, it is tempting to treat everything as if it were ~~a nail~~ a generation problem.

- Seq2seq models (in their current form) are not appropriate for structured prediction
  - Yao & Keller hypothesis: they have access to syntactic information but can't make anything out of it
  - Recall that the early seq2seq parser (Vinyals et al 2014) needed 11M sentences to compare to sota parsers trained on PTB
- On-the-shelf parsers are much better at structural generalisation
  - Weissenhorn & al 2022: concept prediction + graph parsing
  - Liu et al 2021: syntactic module (unlaballed structure) + semantic module (that predicts the nature of semantic links)

COGS and ReCOGS (Wu, Manning, Potts, 2023)

- Some of the difficulty of COGS is due to the format
- The variable-free format is not semantically equivalent to the initial COGS format
  - `need(agent=zebra,xcomp=walk(agent=zebra))`
  - *A zebra needs to walk*
  - *A zebra needs a zebra to walk*

COGS and ReCOGS (Wu, Manning, Potts, 2023)

Variant	Logical Form (LF)
COGS	`zebra(x_1) AND need.agent(x_2,x_1) AND need.xcomp(x_2,x_4) AND walk.agent(x_4,x_1)`
ReCOGS	`zebra(47) ; need(13) AND agent(13,47) AND xcomp(13,48) AND walk(48) AND agent(48,47)`

- remove redundant and useless information (`need.`, `x _`)
- use arbitrary integers for each token (instead of position-based)
- data augmentation:
  - form new examples by concatenating existing exampes
  - word order manipulation (prepose objects + add filler words)
- harmonize treatment of definite and indefinite nouns
- harmonize treatment of proper nouns and common nouns

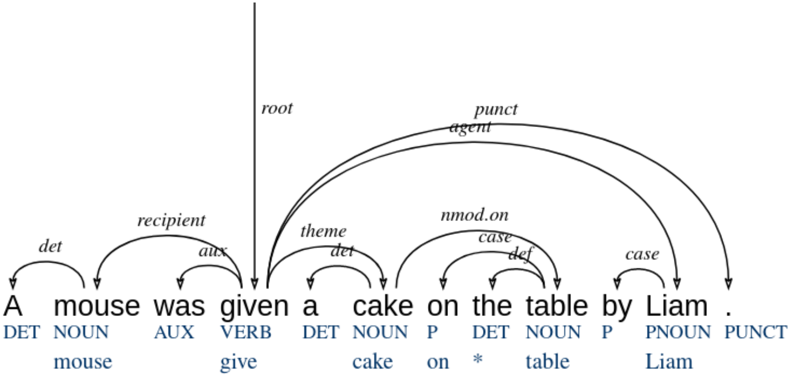
Proposal 1: Reducing COGS to dependency parsing

Motivations:

- COGS / SLOG Semantic representations are **anchored**
  - direct connection between concept / predicate nodes and tokens
- COGS / SLOG graph representations can be transformed to **trees without loss of information**
- parsers are supposedly better at structural generalisations than seq2seq models

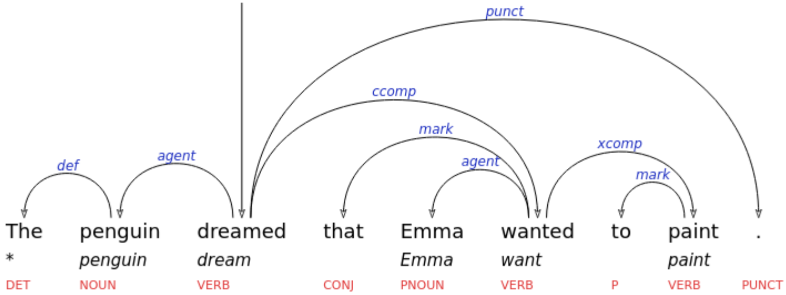
Proposal 1: Reducing COGS to dependency parsing

- `A mouse was given a cake on the table by Liam .`
- `give(recipient=mouse, theme=cake(nmod.on=*table), agent=Liam)`



- use semantic role labels instead of syntactic functions
- borrow Universal Dependencies labels for function words

Proposal 1: Reducing COGS to dependency parsing



- `The penguin dreamed that Emma wanted to paint .`
- `dream(agent=*penguin, ccomp=want(agent=Emma, xcomp=paint(agent=Emma)))`
- Recover the missing arc `(paint, agent, emma)` by finding the agent of the control verb
  - agent of paint is the agent of paints `xcomp` parent

Proposal 1: Reducing COGS to dependency parsing

Work in progress:

- ☒ implement a slog2conllu algorithm
- ☒ implement a conllu2slog algorithm
  - evaluation:
    - transform the full SLOG corpus to conllu
    - transform back to SLOG format
    - eval with exact match
    - 100% reconstruction accuracy on dev and test
    - still some bugs on edge cases for train and gen
- ☒ run dependency parser
  - preliminary experiments this week: exact match = 37%
  - comparisons:
    - Vanilla transformer: 27%
    - T5: 40.6%
    - LLAMA: 40.1%
    - AM-Parser (semantic parser): 70.8%

Invariant language modelling

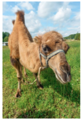
Invariant Language Modeling, EMNLP 2022
Maxime Peyrard, Sarvjeet Ghotra, Martin Josifoski, Vidhan Agarwal, Barun Patra, Dean Carignan, Emre Kiciman, Saurabh Tiwary, Robert West

Background: causal machine learning

- statistical machine learning relies on observed correlations that may be **spurious** or **causal**
  - Underlying assumptions: train and test sets are drawn from the same distribution (IID) and feature the same correlations
  - if there is a distribution shift between train and test settings: poor generalizations

Background: causal machine learning

- statistical machine learning relies on observed correlations that may be **spurious** or **causal**
  - Underlying assumptions: train and test sets are drawn from the same distribution (IID) and feature the same correlations
  - if there is a distribution shift between train and test settings: poor generalizations




Background: causal machine learning

- statistical machine learning relies on observed correlations that may be **spurious** or **causal**
  - Underlying assumptions: train and test sets are drawn from the same distribution (IID) and feature the same correlations
  - if there is a distribution shift between train and test settings: poor generalizations

Causal machine learning:

- Assumes the data was generated from a causality network
- Leverages the knowledge of the causal graph (or tries to discover it)

Background: causal machine learning

- statistical machine learning relies on observed correlations that may be **spurious** or **causal**
  - Underlying assumptions: train and test sets are drawn from the same distribution (IID) and feature the same correlations
  - if there is a distribution shift between train and test settings: poor generalizations

Causal machine learning:

- Assumes the data was generated from a causality network
- Leverages the knowledge of the causal graph (or tries to discover it)
- Motivation 1: **Intepretability** ML: understanding the causal relations in the data leads to explainable decisions
- Motivation 2: **Robustness** to distribution shift

Invariant language modelling (Peyrard et al 2022)

- Assume a collection of datasets (textual corpora), called **environments**
  - each of them has a different distribution (topic, lexical distribution, formality degree, amount of preprocessing / cleaning)
  - . . . and its own biases

Invariant language modelling (Peyrard et al 2022)

- Assume a collection of datasets (textual corpora), called **environments**
  - each of them has a different distribution (topic, lexical distribution, formality degree, amount of preprocessing / cleaning)
  - ... and its own biases
- Goal: **Invariant feature learning**: the language model should learn features that are stable across environments ($\approx$ **causal**) and avoid learning those that are idiosyncratic to some environments (likely to be spurious correlations
- Proposal: a training method for language modelling based on **Invariant Risk Minimization**
- Evaluation on several artificial experimental settings:
  - robustness to noisy training data
  - gender bias mitigaation
  - out-of-domain classification

Environment examples

Noise robustness

- Artificial setting with 2 versions of wikipedia (2 **environments**):
  - text-only: the text of articles is extracted through a classical pipeline
  - full wikipedia html pages
- A typical masked language model will learn to predict html markup syntax
- An invariant language model should ignore html markup because it is absent from the 2nd environment and thus not a feature that is **stable across environment**

Environment examples

Noise robustness

- Artificial setting with 2 versions of wikipedia (2 **environments**):
  - text-only: the text of articles is extracted through a classical pipeline
  - full wikipedia html pages
- A typical masked language model will learn to predict html markup syntax
- An invariant language model should ignore html markup because it is absent from the 2nd environment and thus not a feature that is **stable across environment**

Gender bias

- 2 environments
  - X% wikitext (dataset with strong known gender bias)
  - 100 - X% wikitext with all gendered terms changed
- typical masked language should be unbiased as regards gender iff $X = 50\%$
- invariant language model should be unbiased whatever the value of X

Method: Invariant risk minimization

- Start from a pretrained language models (BERT, distillBERT, Roberta)
- Initialize n language modelling heads parameters w_e, one for each environment e.
  - modelling head $=$ linear $+$ layer norm $+$ linear

Method: Invariant risk minimization

- Start from a pretrained language models (BERT, distillBERT, Roberta)
- Initialize n language modelling heads parameters w_e, one for each environment e.
  - modelling head = linear + layer norm + linear
- Prediction:

$$\hat{y} = \text{softmax}\left(\frac{1}{n}\sum_{e=1}^{n} w_e \circ \phi(x_i)\right)$$

  - average the outputs of each head before softmaxing
  - $\phi(x_i)$ is the BERT representation for token x

Method: Invariant risk minimization

- Start from a pretrained language models (BERT, distillBERT, Roberta)
- Initialize n language modelling heads parameters w_e, one for each environment e.
  - modelling head = linear + layer norm + linear
- Prediction:

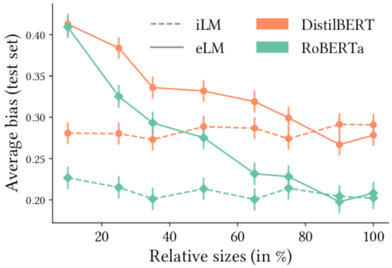
$$\hat{y} = \text{softmax}\left(\frac{1}{n}\sum_{e=1}^{n} w_e \circ \phi(x_i)\right)$$

  - average the outputs of each head before softmaxing
  - $\phi(x_i)$ is the BERT representation for token x
- Training:
  - sample a training example uniformly from any environment e
  - for an example from environment e:
    - compute the (masked LM) loss as above
    - but update only w_e

Results

- baselines:
  - eLM: classical loss, trained on union of environments
  - mtLM: multitask, one head per environment, classical mt training
  - ensLM: ensemble, same as iLM but update every LM head at each training step

1. Noise robustness: iLM $<$ mtLM $<$ ensLM $<$ eLM, eval on perplexity (on clean wikipedia test set)
2. gender bias:
   - measure bias: 1 - entropy of output layer when predicting a gendered term [he, she]
3. domain adaptation (see paper)



Discussion

- Invariant risk minimization: **injecting inductive bias** into model through **environment design**
  - requires a theory / an hypothesis of what are causal / spurious correlations in a dataset
- Outstanding questions:
  - Can Invariant Risk Minimization help a language model learn structure (unsupervisedly)?
  - Can Invariant Risk Minimization help make the type of structural generalizations COGS/SLOG need?

Discussion: syntactic constraints as invariants

- classical data: verb number agreement in English
  - the **keys are**/*is on the table
  - the **keys** to the cabinet **are**/*is on the table
  - Alex's **keys are**/*is on the table
  - the **keys** on the table by the windows **are**/*is big

Discussion: syntactic constraints as invariants

- classical data: verb number agreement in English
  - the **keys are**/*is on the table
  - the **keys** to the cabinet **are**/*is on the table
  - Alex's **keys are**/*is on the table
  - the **keys** on the table by the windows **are**/*is big
- Given a toy datasets, we can make several generalization:
  - **Structural generalization**: verb agrees with its subject
    - invariant: syntactic constraints
  - **Linear generalization**: verb agrees with last preceding noun
    - spurious correlation, but a heuristics that works very frequently!!
- Linzen et al 2016, Gulordava et al 2018: LSTM language models learn subject-verb agreement almost as well as speakers
- Can they learn subject-agreement with less data when trained with Invariant Risk Minimization?