EE-608: Deep Learning For Natural Language Processing:

Future Challenges

James Henderson

Idiap Research Institute

DLNLP, Lecture 12

Announcements

- Project presentations next week, during both lecture and exercises sessions. Let us know soon if you have time constraints.
- We encourage everyone to do their presentations in-person, but doing it by Zoom is allowed, including hybrid presentations. Please organise this with your team members.
- Watching the presentations of other groups via Zoom will be possible.

Outline

Outline

The Unstoppable Rise of Computational Linguistics in Deep Learning

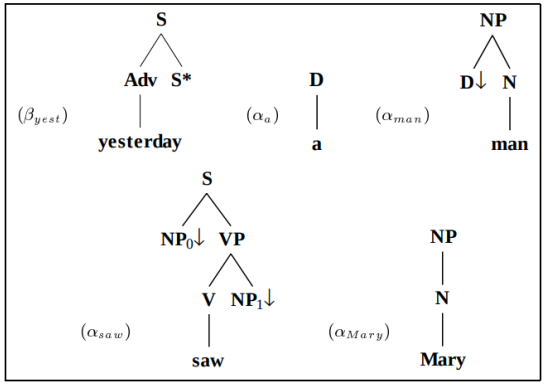
This section is from the ACL 2020 theme paper:

James Henderson. 2020. “The Unstoppable Rise of Computational Linguistics in Deep Learning”. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6294–6306, Online.

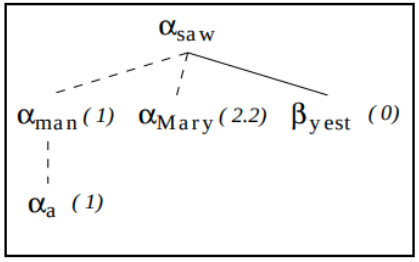
The Nature of Language

Grammar Formalisms are designed to capture the generalisations found in natural languages.

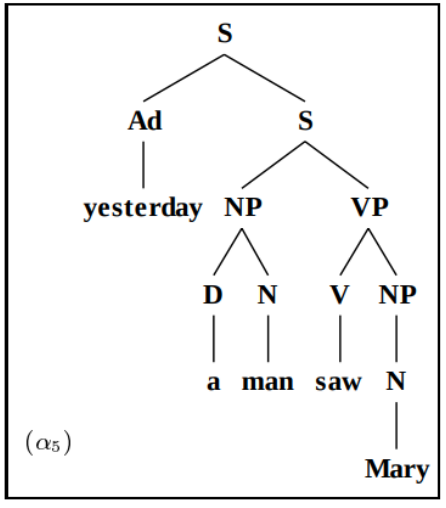
- Multiple levels of representation
- each with multiple entities
- each with multiple properties
- and relations between entities
- Rules specify the allowable structures
- Grammar formalisms specify the allowable rules



TAG Grammar



Derivation structure

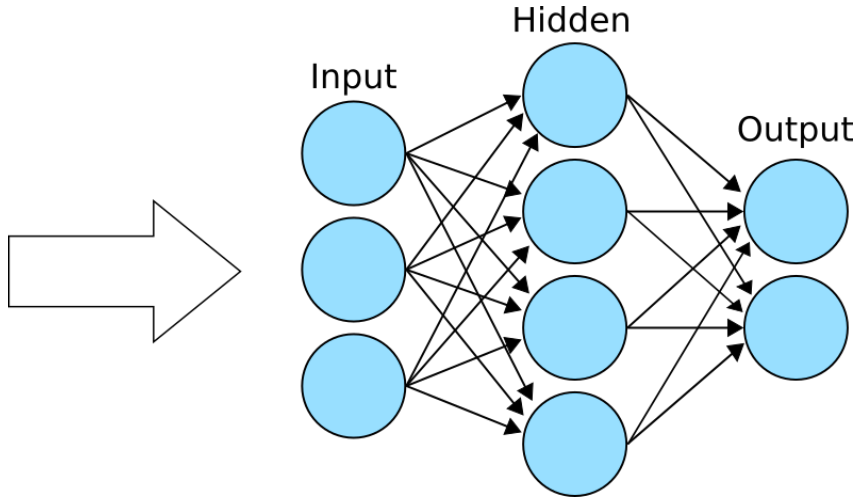


Linguistic structure

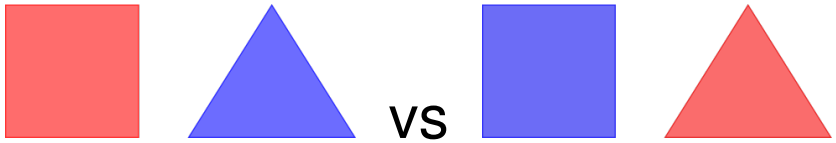
Connectionism

Brains have distributed representations and learning.

- Vector spaces formalise distributed representations
- Backpropagation learning is astoundingly effective
- Multi-layer and recurrent neural networks are powerful approximators

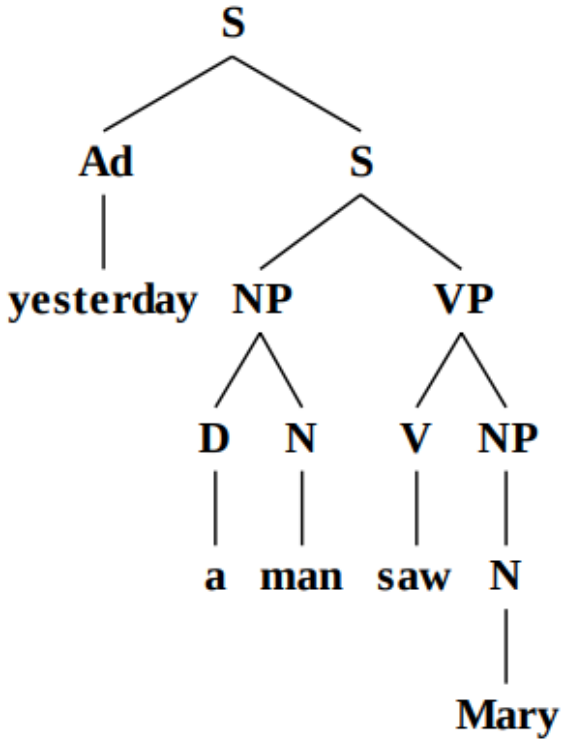


Variable Binding



Questioning the adequacy of vector spaces.

- Our representations segment the world into entities
- Variable binding: How are multiple entities represented in the brain?
- Systematicity: How can we learn rules which generalise across entities?



Overview

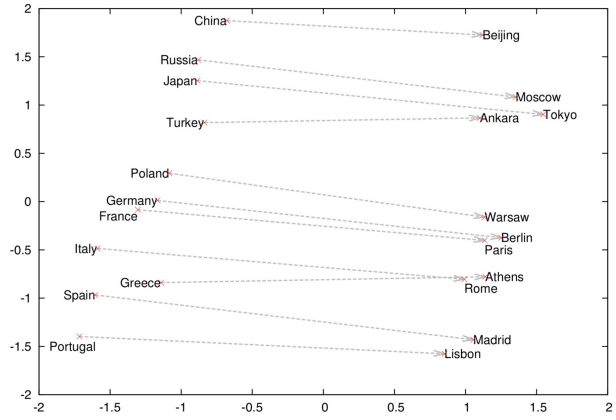
Our understanding of the nature of language from computational linguistics has fundamentally influenced deep learning architectures.

1. Inducing Features of Entities
2. Inducing Relations between Entities
3. Inducing Entities?

Early Neural Network Models of Language

Neural networks were successfully integrated in classical natural language understanding models

- Learned vectors replace categorical labels
  - Word embeddings
  - Contextualised token embeddings
  - Parse history embeddings

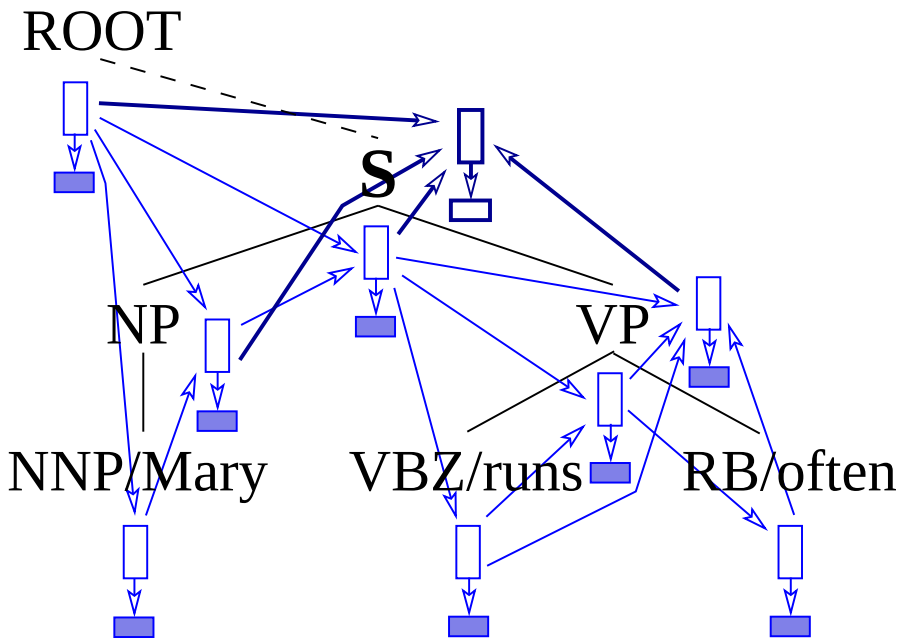


- Neural probability estimation replaces feature engineering
  - Subtree scores
  - Parser action scores

Modelling Derivation Structures

But successful neural approaches kept the same structures

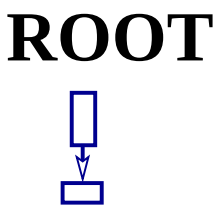
- Linguistic structures reflect regularities
- Derivation structures reflect the linguistic structure
- Hand-coded neural network model structures reflect the derivation structure



Modelling Derivation Structures

But successful neural approaches kept the same structures

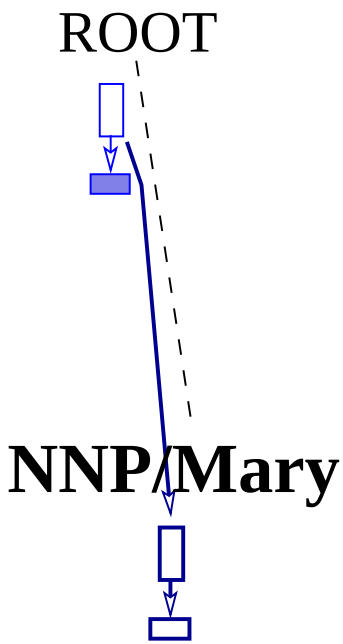
- Linguistic structures reflect regularities
- Derivation structures reflect the linguistic structure
- Hand-coded neural network model structures reflect the derivation structure



Modelling Derivation Structures

But successful neural approaches kept the same structures

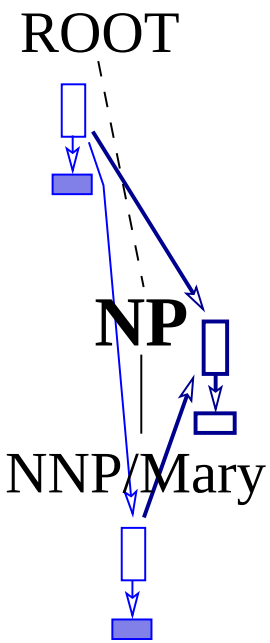
- Linguistic structures reflect regularities
- Derivation structures reflect the linguistic structure
- Hand-coded neural network model structures reflect the derivation structure



Modelling Derivation Structures

But successful neural approaches kept the same structures

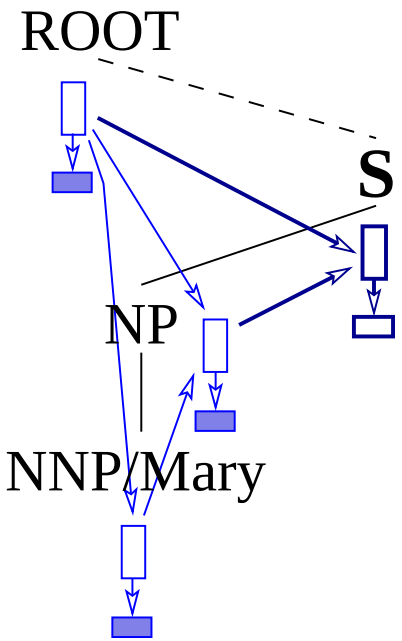
- Linguistic structures reflect regularities
- Derivation structures reflect the linguistic structure
- Hand-coded neural network model structures reflect the derivation structure



Modelling Derivation Structures

But successful neural approaches kept the same structures

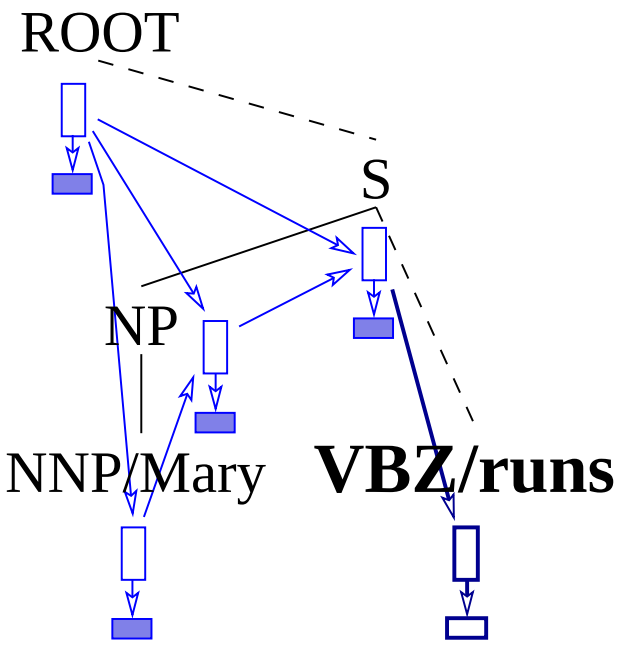
- Linguistic structures reflect regularities
- Derivation structures reflect the linguistic structure
- Hand-coded neural network model structures reflect the derivation structure



Modelling Derivation Structures

But successful neural approaches kept the same structures

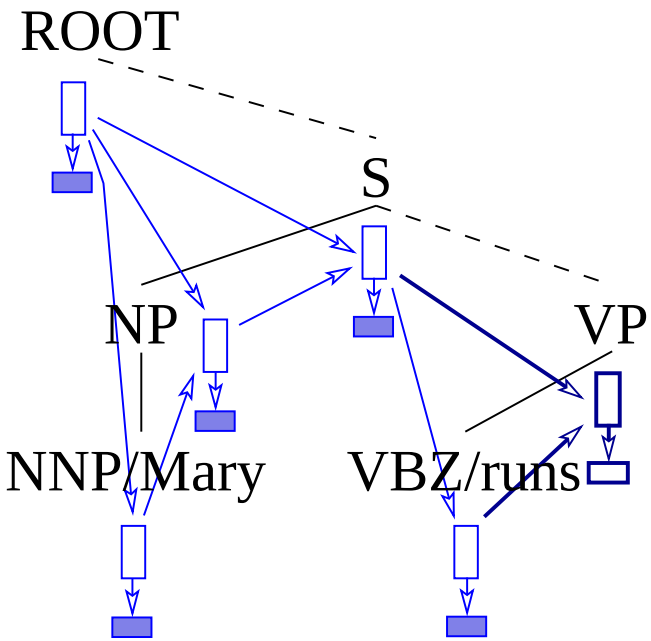
- Linguistic structures reflect regularities
- Derivation structures reflect the linguistic structure
- Hand-coded neural network model structures reflect the derivation structure



Modelling Derivation Structures

But successful neural approaches kept the same structures

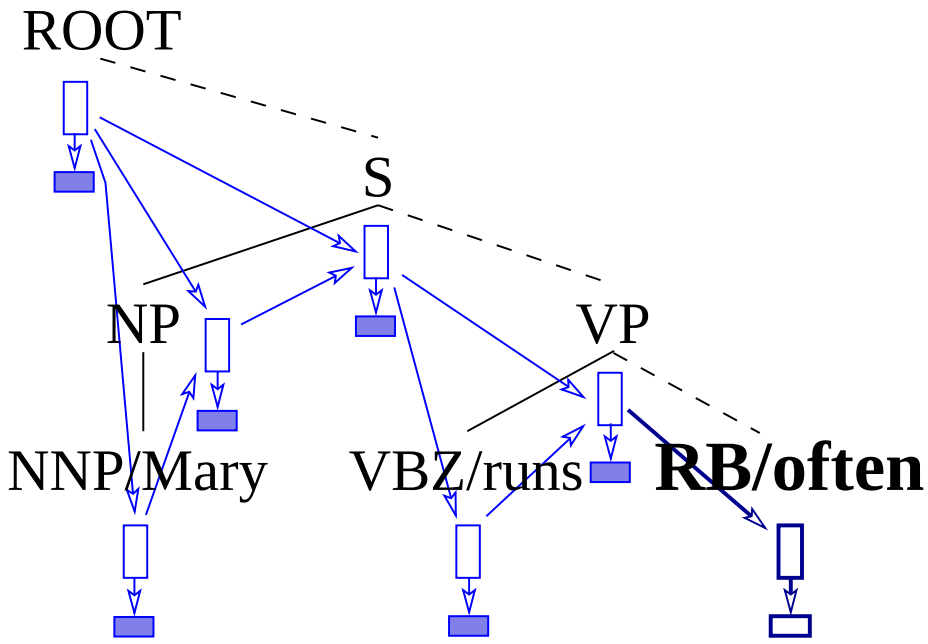
- Linguistic structures reflect regularities
- Derivation structures reflect the linguistic structure
- Hand-coded neural network model structures reflect the derivation structure



Modelling Derivation Structures

But successful neural approaches kept the same structures

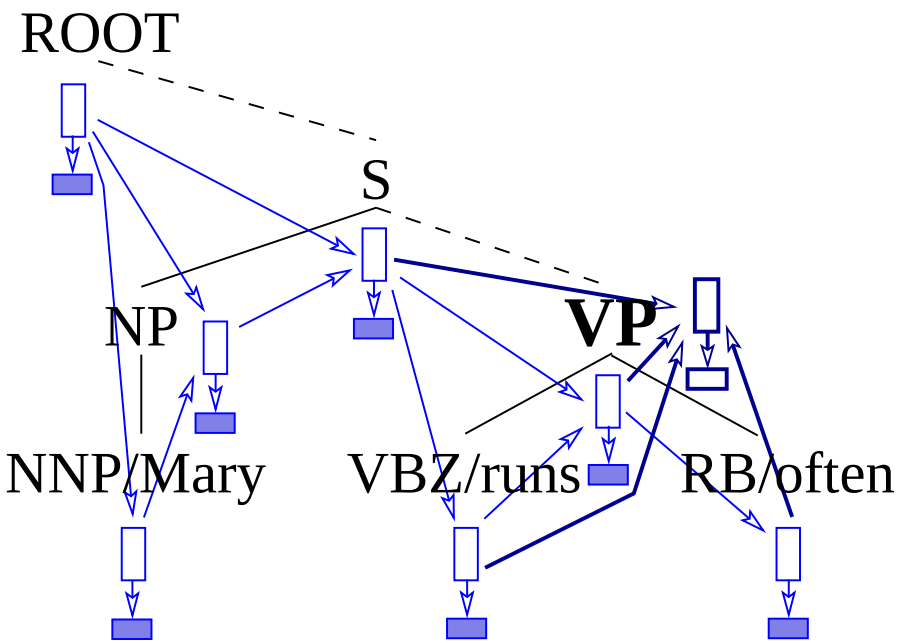
- Linguistic structures reflect regularities
- Derivation structures reflect the linguistic structure
- Hand-coded neural network model structures reflect the derivation structure



Modelling Derivation Structures

But successful neural approaches kept the same structures

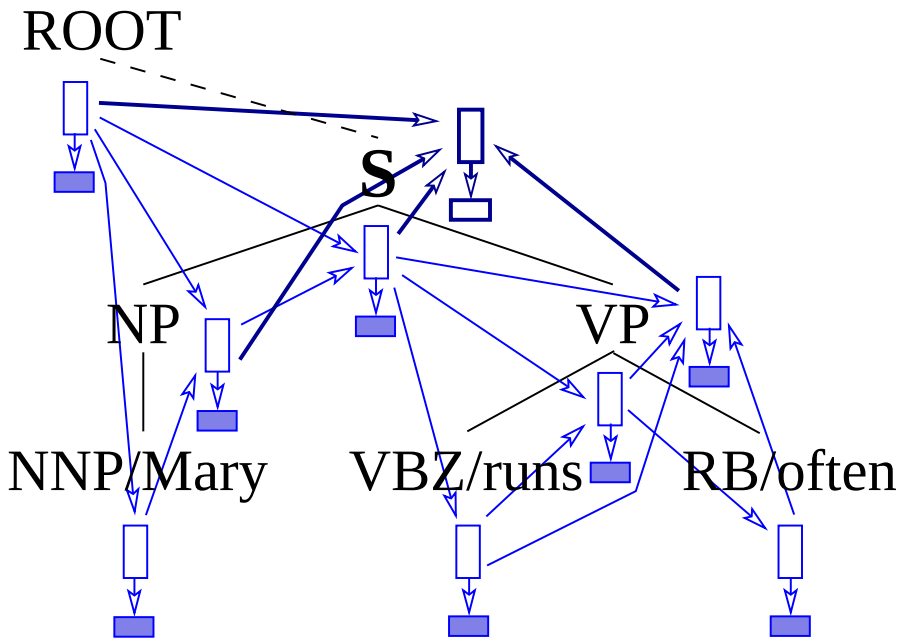
- Linguistic structures reflect regularities
- Derivation structures reflect the linguistic structure
- Hand-coded neural network model structures reflect the derivation structure



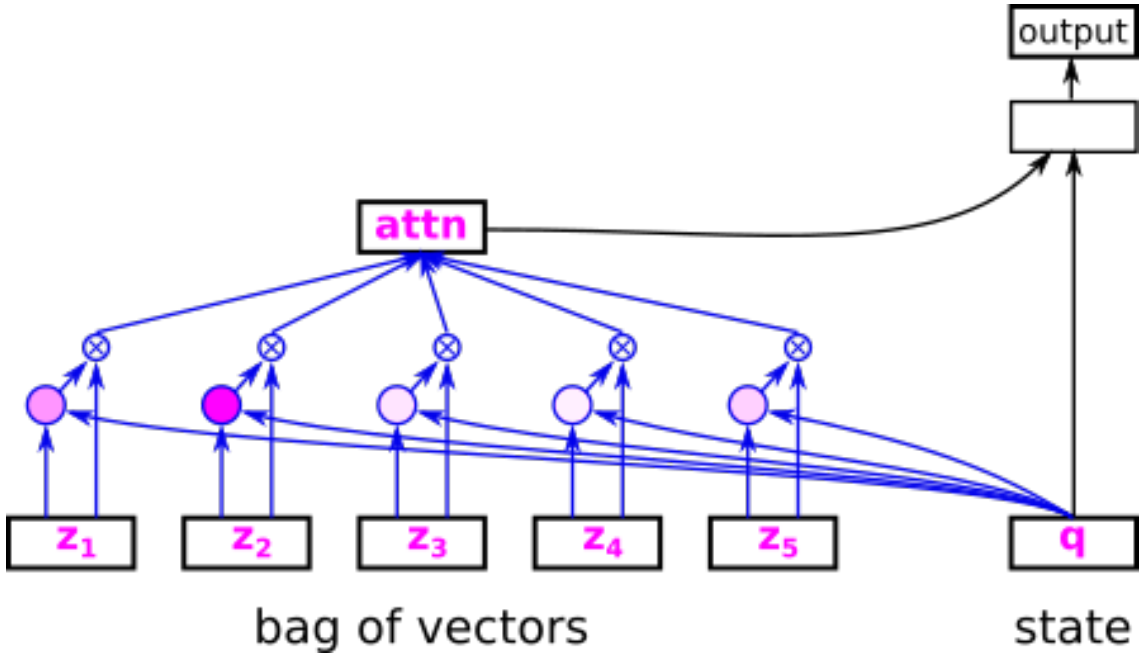
Modelling Derivation Structures

But successful neural approaches kept the same structures

- Linguistic structures reflect regularities
- Derivation structures reflect the linguistic structure
- Hand-coded neural network model structures reflect the derivation structure



Attention Induces Structure



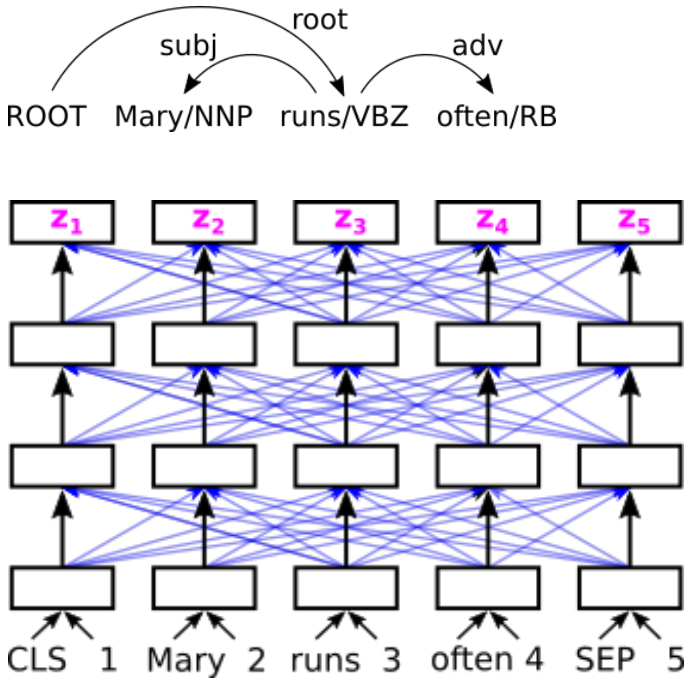
Attention-based models are induced structure models, not sequence models

- Attention weights determine the model structure
- Model structure reflects linguistic structure
- Attention treats a sequence as a bag-of-vectors

Transformers and Variable Binding

Transformers have variable binding.

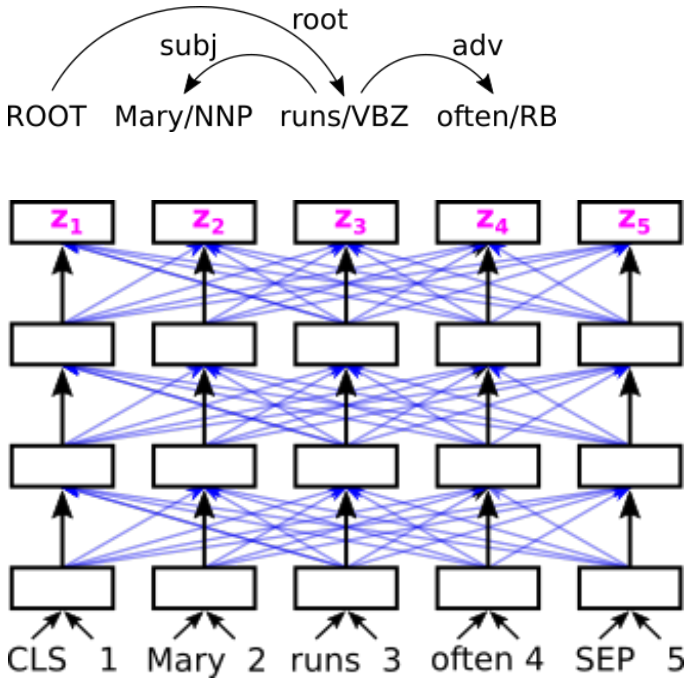
- Each “position” is an entity
- Each vector embeds the properties of its entity
- Each pair of vectors embeds the relations between their entities



Transformers and Systematicity

Transformers have systematicity.

- Rules are parameterised in terms of an entity's properties and relations
- Parameters are shared across entities
- Rules learned for one entity inherently generalise to other entities



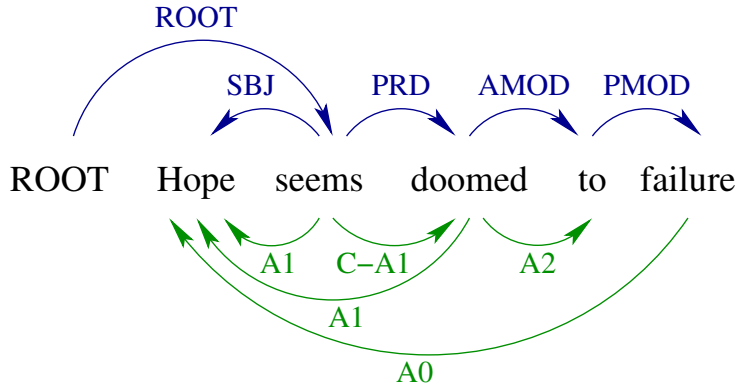
Future Directions

What is left to learn?

- How do we learn the set of entities?

  - Currently, the set of positions are pre-defined, with hand-coding (e.g. words), or preprocessing (e.g. BPE)

- How do we learn the set of levels?
  - Linguistic levels seem to be language universal
  - Are they innate?
  - Can they be learned at all?



Summary

Our understanding of the nature of language from computational linguistics has fundamentally influenced deep learning architectures.

- Vector spaces and sequence-to-sequence models are **not adequate**
- Attention-based models **induce structured representations** over entities
- Transformers use bag-of-vector representations (variable binding)
- Transformers inherently generalise across entities (systematicity)

But levels of representation and their entities are predefined. Can they be learned from data?

Outline

Limitations of current LLMs

- LLMs excel at many tasks but still struggle on low-resource languages, systematic generalisation, and abstraction.

- Need: curated diagnostic tasks that probe representation of objects, structure-dependency and systematicity.

Slide from Paola Merlo

Do LLMs know the grammar and the meaning of the language they seem to use and generate so well?

The truck was loaded with hay. The hay was loaded onto the truck.

Locative Instrumental Instrumental Locative

- Verb alternations require object identification, attribute assignment and systematic correspondence.
- Can current LLMs perform these linguistic operations?

Slide from Paola Merlo

More problems and underlying rules

Examples of verb alternations (Levin 1993) (Italian, English, Hebrew)

The man rolled the dice/ The dice rolled (manner of motion)

The cook burnt the cake/ The cake burnt (Change of state)

The artist paints the portrait/ The artist paints (object-drop)

Example of agreement (French, Italian, English, Rumanian)

L'ordinateur avec le programme est en panne.

Les ordinateurs avec le programme sont en panne.

L'ordinateur avec le programme de l'expérience est en panne.

Les ordinateurs avec le programme de l'expérience sont en panne.

Slide from Paola Merlo

The architectural and data-centric approach

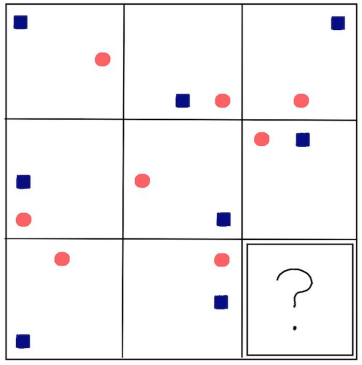
- We develop curated synthetic data on a large scale. We develop **data and diagnostic probes** that help us understand their current generalisation abilities.

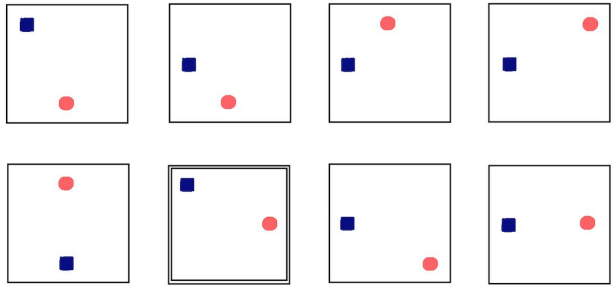
Merlo 2023 Arvix
An et al 2023 EACL, Samo et al 2023, Findings EMNLP, Merlo 2023, Findings Emnlp,

Slide from Paola Merlo

The task: BLMs

RPMs





BLMs

CONTEXT				
1	NP-sg	PP1-sg		VP-sg
2	NP-pl	PP1-sg		VP-pl
3	NP-sg	PP1-pl		VP-sg
4	NP-pl	PP1-pl		VP-pl
5	NP-sg	PP1-sg	PP2-sg	VP-sg
6	NP-pl	PP1-sg	PP2-sg	VP-pl
7	NP-sg	PP1-pl	PP2-sg	VP-sg
8	???			

ANSWERS					
1	NP-pl	PP1-pl	PP2-sg	VP-pl	**Corr**
2	NP-pl	PP1-pl	et PP2-sg	VP-pl	Coord
3	NP-pl	PP1-pl		VP-pl	WNA
4	NP-pl	PP1-sg	PP1-sg	VP-pl	WN1
5	NP-pl	PP1-pl	PP2-pl	VP-pl	WN2
6	NP-pl	PP1-pl	PP2-pl	VP-sg	AEV
7	NP-pl	PP1-sg	PP2-pl	VP-sg	AEN1
8	NP-pl	PP1-pl	PP2-sg	VP-sg	AEN2

Slide from Paola Merlo

The task: BLMs

	CONTEXT			
1	NP-sg	PP1-sg		VP-sg
2	NP-pl	PP1-sg		VP-pl
3	NP-sg	PP1-pl		VP-sg
4	NP-pl	PP1-pl		VP-pl
5	NP-sg	PP1-sg	PP2-sg	VP-sg
6	NP-pl	PP1-sg	PP2-sg	VP-pl
7	NP-sg	PP1-pl	PP2-sg	VP-sg
8	???			

	ANSWERS				
1	NP-pl	PP1-pl	PP2-sg	VP-pl	**Corr**
2	NP-pl	PP1-pl	et PP2-sg	VP-pl	Coord
3	NP-pl	PP1-pl		VP-pl	WNA
4	NP-pl	PP1-sg	PP1-sg	VP-pl	WN1
5	NP-pl	PP1-pl	PP2-pl	VP-pl	WN2
6	NP-pl	PP1-pl	PP2-pl	VP-sg	AEV
7	NP-pl	PP1-sg	PP2-pl	VP-sg	AEN1
8	NP-pl	PP1-pl	PP2-sg	VP-sg	AEN2

What kind of linguistic objects the LLM actually learn and manipulate?
Are they tokens?
Does the LLM discover something more linguistically justified like phrases and constituents?

Are the inner representations of NNs information about chunks mapped to semantic roles? Are long-distance dependencies in agreement represented?

Are the objects and attributes represented by higher-level abstractions or based on low-level cues?

Do the abstractions hold across languages?

Slide from Paola Merlo

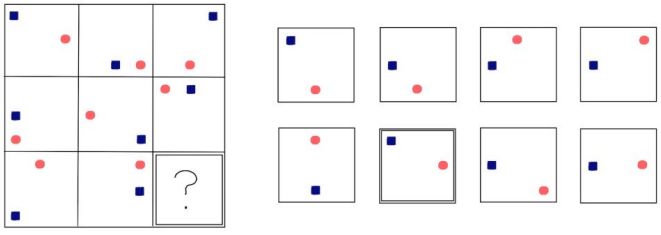
Blackbird Language Matrices: A Framework to Investigate the Linguistic Competence of LLMs

Paola Merlo
Computational Learning and Computational Linguistics group

www.idiap.ch

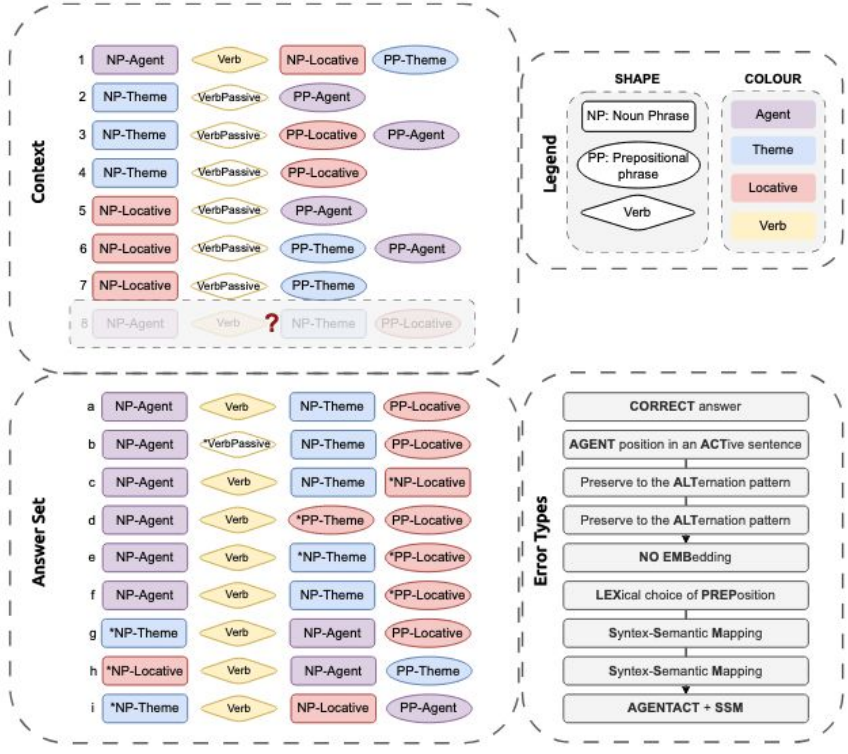
Slide from Paola Merlo

The BLM task



Beyond performance: human-like generalisation in NN

- Rule-based generalisation
- With little data
- Across languages

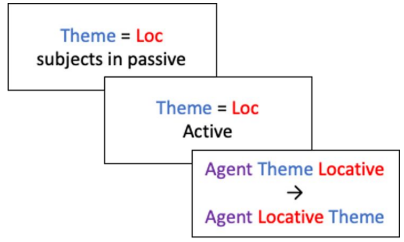


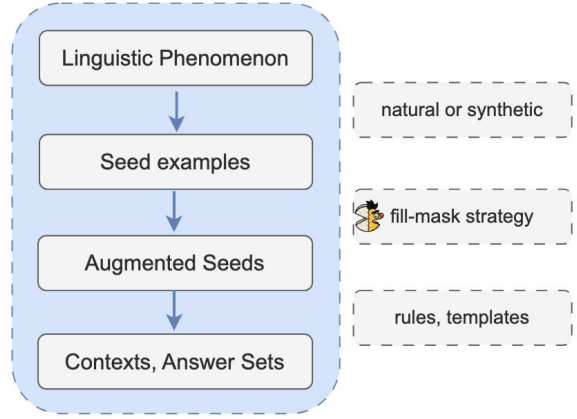
Aixiu An, Chunyang Jiang, Maria A. Rodriguez, Vivi Nastase, and Paola Merlo. (2023). *BLM-AgrF: A New French Benchmark to Investigate Generalization of Agreement in Neural Networks.* In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, p. 1363–1374, Dubrovnik, Croatia.

Slide from Paola Merlo

Creating matrices and datasets

	Context			
1	NP-Agent	Verb	NP-Loc	PP-Theme
2	NP-Theme	VerbPass	PP-Agent	
3	NP-Theme	VerbPass	PP-Loc	PP-Agent
4	NP-Theme	VerbPass	PP-Loc	
5	NP-Loc	VerbPass	PP-Agent	
6	NP-Loc	VerbPass	PP-Theme	PP-Agent
7	NP-Loc	VerbPass	PP-Theme	
?8	NP-Agent	Verb	NP-Theme	PP-Loc





Agreement

English French

Italian Romanian

Spray/Load

She loads the truck with hay

She loads hay into the truck

Causatives

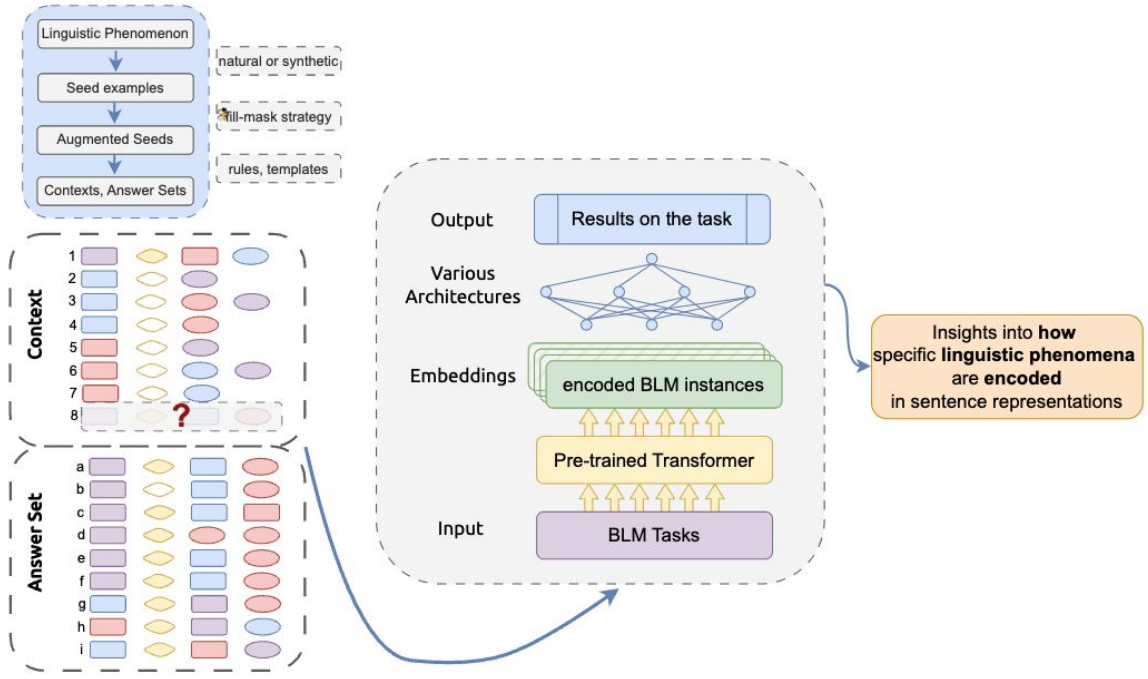
Mary opened the door

The door opened.

G. Samo, V. Nastase, C. Jiang, and P. Merlo. 2023. *BLM-s/lE: A structured dataset of English spray-load verb alternations for testing generalization in LLMs.* In: Findings of the Association for Computational Linguistics: EMNLP 2023, pages 12276–12287, Singapore.

Slide from Paola Merlo

The framework pipeline



V. Nastase and P. Merlo. (2024). *Tracking linguistic information in transformer-based sentence embeddings through targeted sparsification.* In: Proceedings of the 9th Workshop on Representation Learning for NLP (RepL4NLP-2024), pages 203–214, Bangkok, Thailand.

Slide from Paola Merlo

Experiments

Can BLMs be solved by current models? At what accuracy?

(benchmarks)

Do sentence embeddings encode chunk/constituent structure?

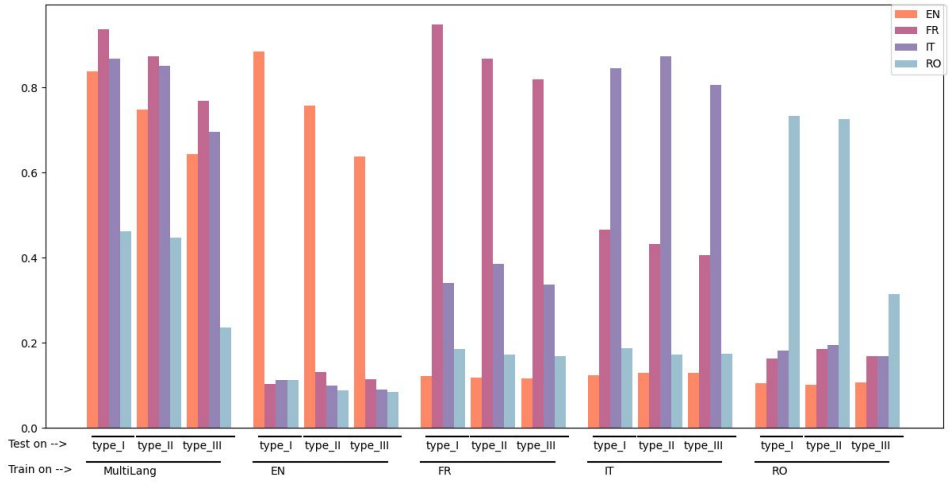
Can compressed representations reveal systematic structure across sentences?

Slide from Paola Merlo

The linguistic results

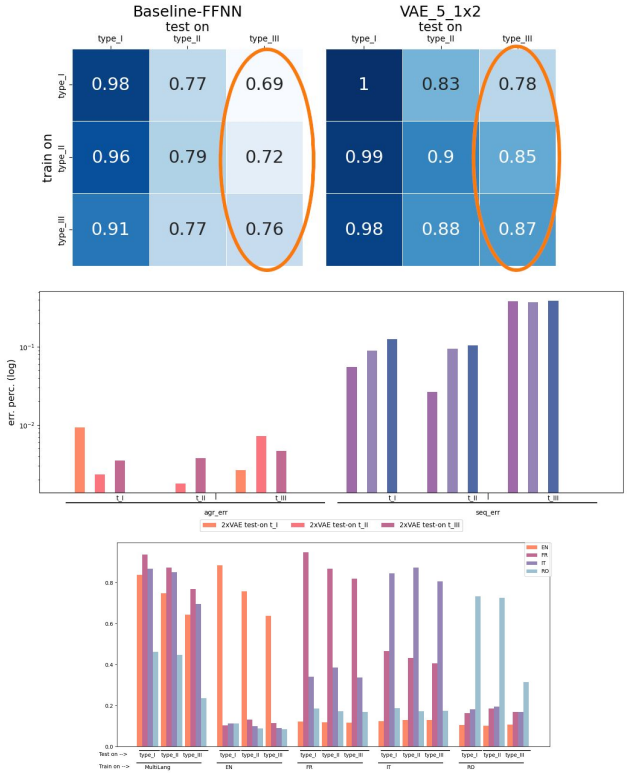
No cross-lingual abstraction

Nastase et al 2024, Clic-it.



Slide from Paola Merlo

Results



The task	Good overall performance, sensitive to lexicalisation
Error analysis	Good on language, less well on sequence reasoning
Across languages	No cross-lingual abstraction.

Slide from Paola Merlo

Outline

Compositional Representations and Systematic Generalization

- Systematicity: The ability to produce/understand some sentences is intrinsically connected to the ability to produce / understand certain others. This means there is a “definite and predictable pattern among the sentences we understand”
- E.g. any speaker that understands the sentence “John loves Mary” should be able to understand “Mary loves John”.

Source: [Stanford Encyclopedia of Philosophy]

Slide from Shikhar Murty

Compositional Representations and Systematic Generalization

Compositionality: closely related to the idea of systematicity is the principle of compositionality.

Rough Definition:

"The meaning of an expression is a function of the meaning of its parts"

More concrete definition (Montague):

A homomorphism from syntax (structure) to semantics (meaning). That is, meaning of the whole is a function of immediate constituents (as determined by syntax)

Slide from Shikhar Murty

Are human languages compositional?

Brown Cow = *Brown objects ∩ Cows*

Red Rabbit = *Red objects ∩ Rabbits*

Kicked the Ball = *Kicked(Ball, Agent)*

Slide from Shikhar Murty

Are human languages compositional?

Brown Cow = *Brown objects ∩ Cows*

Red Rabbit = *Red objects ∩ Rabbits*

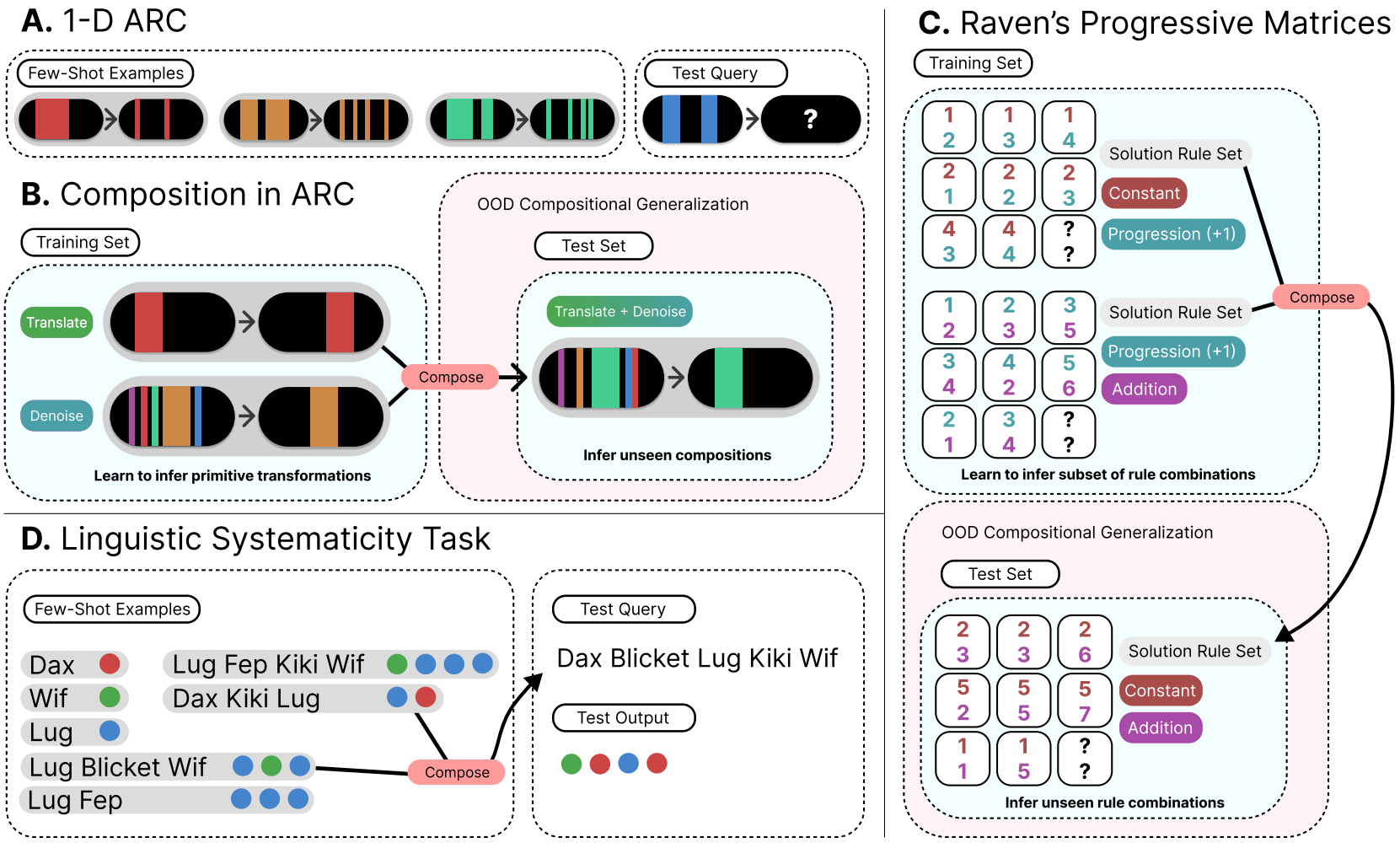
Kicked the Ball = *Kicked(Ball, Agent)*

Red Herring ≠ *Red things ∩ Herring*

Kicked the bucket ≠ *Kicked(Bucket, Agent)*

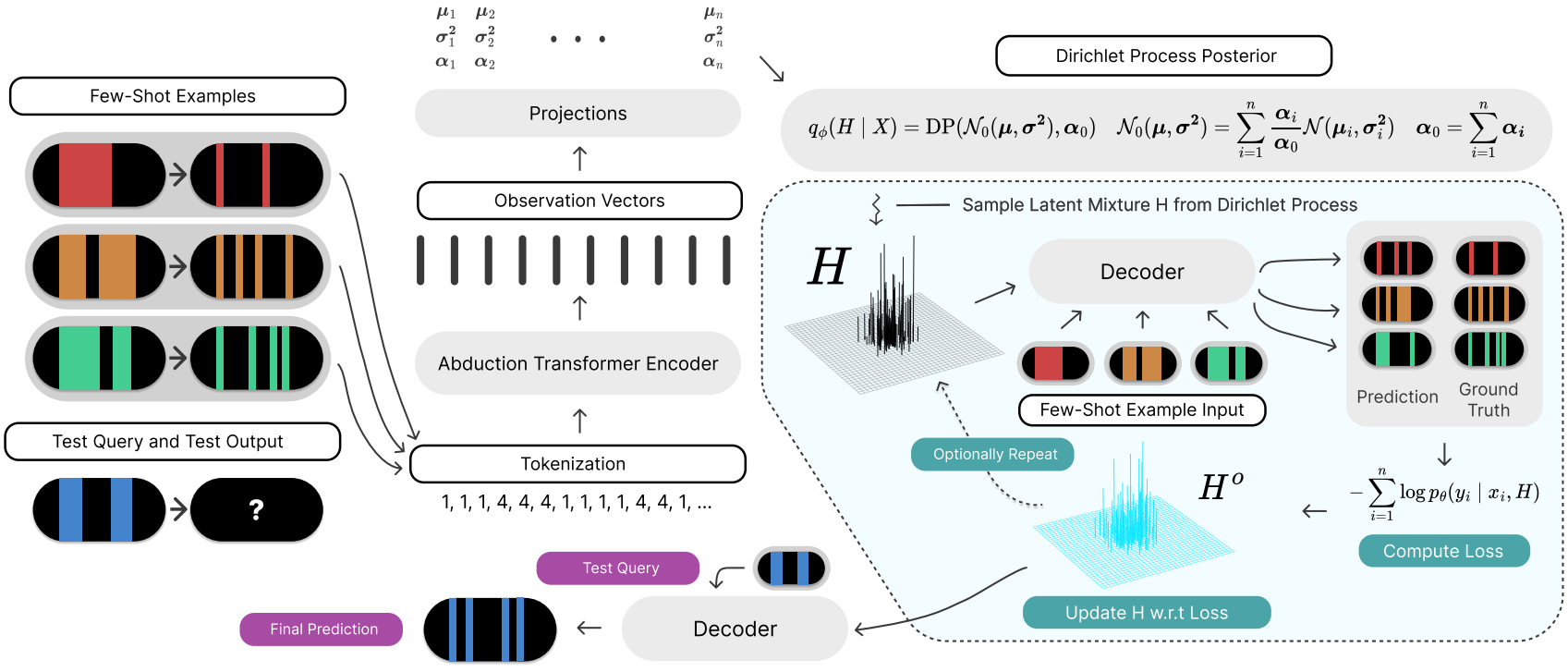
Slide from Shikhar Murty

Compositional Generalisation Tasks



Test sets contain unseen compositions of seen primitives.

Abduction Transformer Model



Search for a latent program which solves the few-shot examples.

- An encoder tries to guess the latent program
- Gradient search fine-tunes the latent program
- Uses a nonparametric latent space to generalise from simple to complex programs

Outline

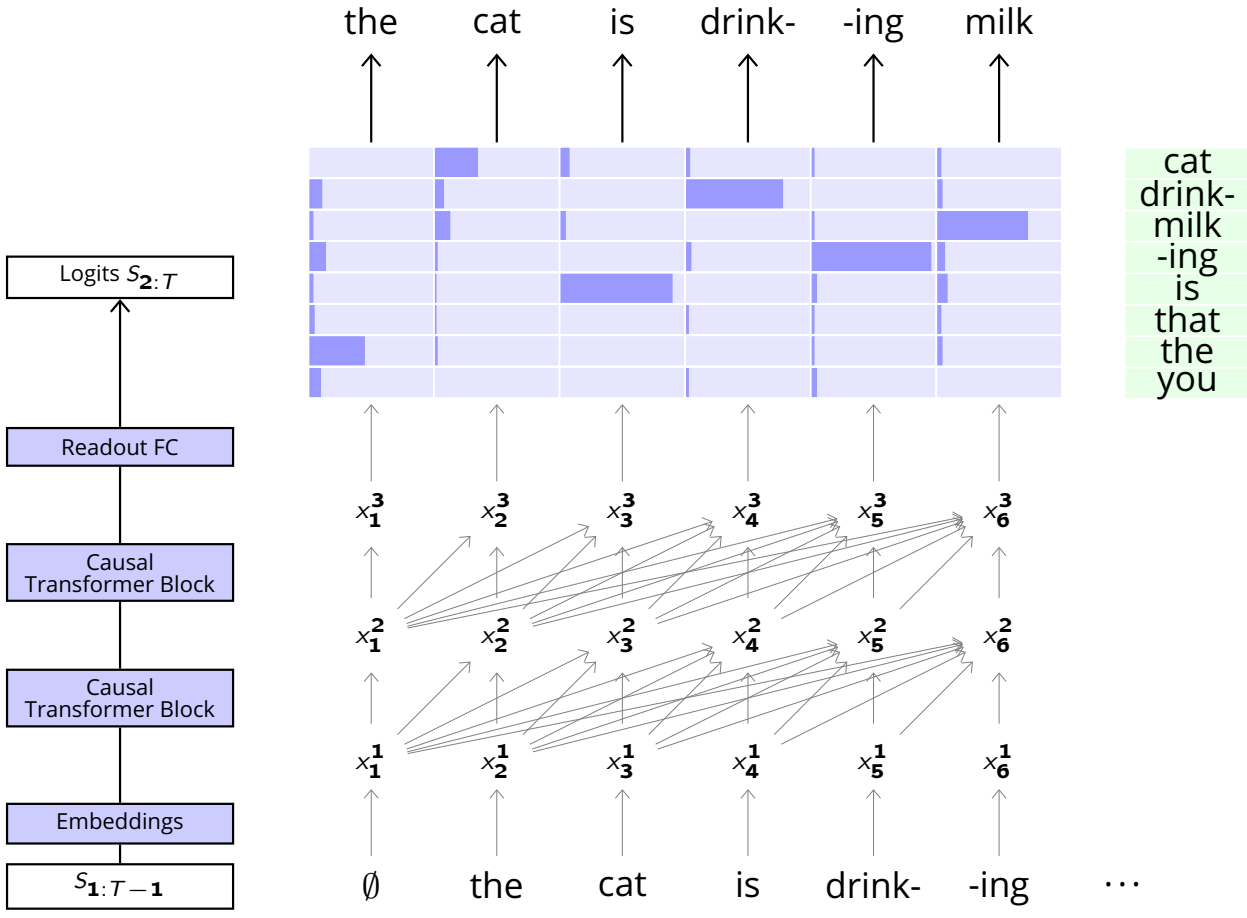
Selected slides from (arXiv 2025):

The Free Transformer

François Fleuret

Meta

Slide from Francois Fleuret



Slide from Francois Fleuret

Such a model, trained only to generate text, can already be put to use for instance for classification.

```
I: water boils at 100 degrees, O: physics.  I: the square root of two is
irrational, O: mathematics.  I: the set of prime numbers is infinite, O:
mathematics.  I: gravity is proportional to the mass, O:
```


Slide from Francois Fleuret

Such a model, trained only to generate text, can already be put to use for instance for classification.

```
I: water boils at 100 degrees, O: physics.  I: the square root of two is
irrational, O: mathematics.  I: the set of prime numbers is infinite, O:
mathematics.  I: gravity is proportional to the mass, O: physics.
```


Slide from Francois Fleuret

Such a model, trained only to generate text, can already be put to use for instance for classification.

```
I: water boils at 100 degrees, O: physics.  I: the square root of two is
irrational, O: mathematics.  I: the set of prime numbers is infinite, O:
mathematics.  I: gravity is proportional to the mass, O: physics.
```

```
I: water boils at 100 degrees, O: physics.  I: the square root of two is
irrational, O: mathematics.  I: the set of prime numbers is infinite, O:
mathematics.  I: squares are rectangles, O:
```


Slide from Francois Fleuret

Such a model, trained only to generate text, can already be put to use for instance for classification.

```
I: water boils at 100 degrees, O: physics.  I: the square root of two is
irrational, O: mathematics.  I: the set of prime numbers is infinite, O:
mathematics.  I: gravity is proportional to the mass, O: physics.
```

```
I: water boils at 100 degrees, O: physics.  I: the square root of two is
irrational, O: mathematics.  I: the set of prime numbers is infinite, O:
mathematics.  I: squares are rectangles, O: mathematics.
```


Slide from Francois Fleuret

Such a model, trained only to generate text, can already be put to use for instance for classification.

```
I: water boils at 100 degrees, O: physics.  I: the square root of two is
irrational, O: mathematics.  I: the set of prime numbers is infinite, O:
mathematics.  I: gravity is proportional to the mass, O: physics.
```

```
I: water boils at 100 degrees, O: physics.  I: the square root of two is
irrational, O: mathematics.  I: the set of prime numbers is infinite, O:
mathematics.  I: squares are rectangles, O: mathematics.
```

```
I: I love apples, O: positive.  I: music is my passion, O: positive.  I: my
job is boring, O: negative.  I: frozen pizzas are awesome, O:
```


Slide from Francois Fleuret

Such a model, trained only to generate text, can already be put to use for instance for classification.

```
I: water boils at 100 degrees, O: physics.  I: the square root of two is
irrational, O: mathematics.  I: the set of prime numbers is infinite, O:
mathematics.  I: gravity is proportional to the mass, O: physics.
```

```
I: water boils at 100 degrees, O: physics.  I: the square root of two is
irrational, O: mathematics.  I: the set of prime numbers is infinite, O:
mathematics.  I: squares are rectangles, O: mathematics.
```

```
I: I love apples, O: positive.  I: music is my passion, O: positive.  I: my
job is boring, O: negative.  I: frozen pizzas are awesome, O: positive.
```


Slide from Francois Fleuret

Such a model, trained only to generate text, can already be put to use for instance for classification.

```
I: water boils at 100 degrees, O: physics.  I: the square root of two is
irrational, O: mathematics.  I: the set of prime numbers is infinite, O:
mathematics.  I: gravity is proportional to the mass, O: physics.
```

```
I: water boils at 100 degrees, O: physics.  I: the square root of two is
irrational, O: mathematics.  I: the set of prime numbers is infinite, O:
mathematics.  I: squares are rectangles, O: mathematics.
```

```
I: I love apples, O: positive.  I: music is my passion, O: positive.  I: my
job is boring, O: negative.  I: frozen pizzas are awesome, O: positive.
```

```
I: I love apples, O: positive.  I: music is my passion, O: positive.  I:
my job is boring, O: negative.  I: frozen pizzas taste like cardboard, O:
```


Slide from Francois Fleuret

Such a model, trained only to generate text, can already be put to use for instance for classification.

```
I: water boils at 100 degrees, O: physics.  I: the square root of two is
irrational, O: mathematics.  I: the set of prime numbers is infinite, O:
mathematics.  I: gravity is proportional to the mass, O: physics.
```

```
I: water boils at 100 degrees, O: physics.  I: the square root of two is
irrational, O: mathematics.  I: the set of prime numbers is infinite, O:
mathematics.  I: squares are rectangles, O: mathematics.
```

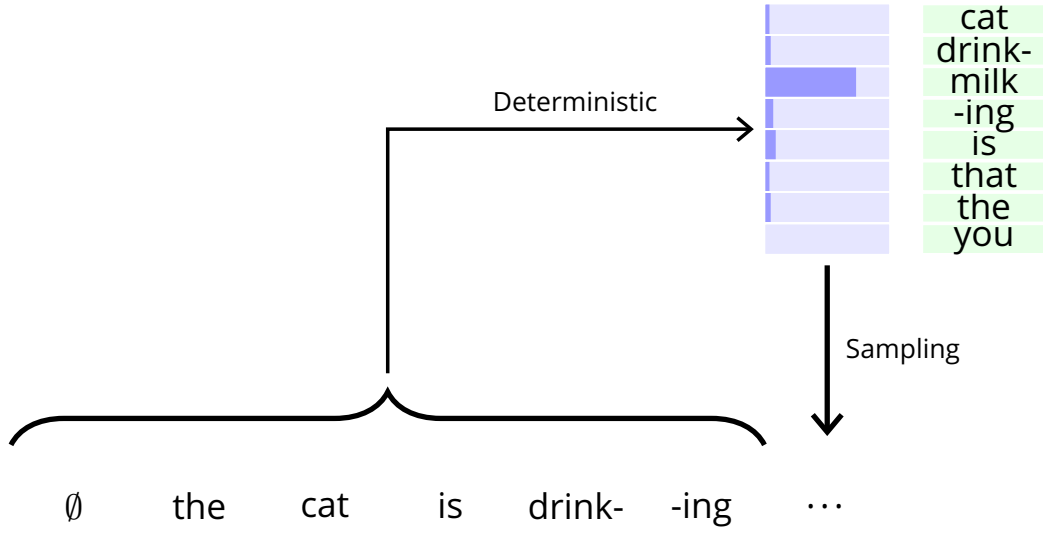
```
I: I love apples, O: positive.  I: music is my passion, O: positive.  I: my
job is boring, O: negative.  I: frozen pizzas are awesome, O: positive.
```

```
I: I love apples, O: positive.  I: music is my passion, O: positive.  I:
my job is boring, O: negative.  I: frozen pizzas taste like cardboard, O:
negative.
```


Slide from Francois Fleuret

The problem with autoregression

Slide from Francois Fleuret



The only sampling in a Transformer are the tokens.

Slide from Francois Fleuret

Consider a corpus of positive and negative online reviews.

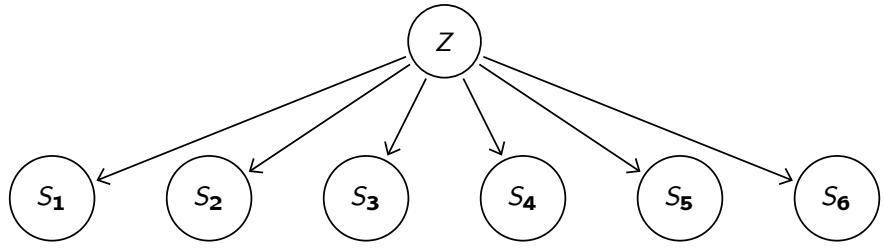
$$Z \sim \mathcal{U}\left(\{-1, 1\}\right), \; S \sim \mu_Z.$$

An AR model has no way of implementing this factorized distribution, it cannot “decide” beforehand what type of review to generate.

It would estimate on the fly the negativity / positivity of what it has written so far to be consistent.

Slide from Francois Fleuret

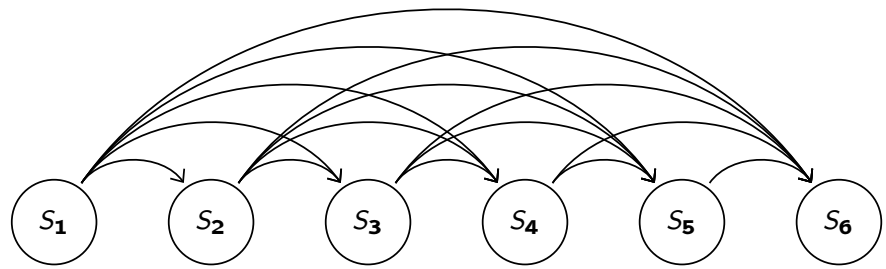
The “true model” of a joint distribution, with latents, is usually simpler and more robust than the “latentless” AR model.



$$\mathrm{P}(S_{t+1} = 1 \mid Z, S_1, \ldots, S_t) = (1 - \epsilon)Z + \epsilon(1 - Z)$$

Slide from Francois Fleuret

The “true model” of a joint distribution, with latents, is usually simpler and more robust than the “latentless” AR model.



$$\mathrm{P}(X_{t+1} = 1 \mid X_1 = x_1, \ldots, X_t = x_t) = \frac{\left(\frac{\epsilon}{1-\epsilon}\right)^{\sum_{s=1}^{t} x_s} (1-\epsilon)^t \epsilon + \left(\frac{1-\epsilon}{\epsilon}\right)^{\sum_{s=1}^{t} x_s} \epsilon^t (1-\epsilon)}{\left(\frac{\epsilon}{1-\epsilon}\right)^{\sum_{s=1}^{t} x_s} (1-\epsilon)^t + \left(\frac{1-\epsilon}{\epsilon}\right)^{\sum_{s=1}^{t} x_s} \epsilon^t}.$$

Slide from Francois Fleuret

We could prefix every training sample with a token Z indicating if the review is positive or not.

Slide from Francois Fleuret

We could prefix every training sample with a token Z indicating if the review is positive or not.

Reasoning post-training tries to address it unsupervised, however:

+ it requires a trained model,
+ current approaches cast it as reinforcement learning,
+ the conditionning variable are discrete tokens,
+ it cannot deal with stochastic responses.

Slide from Francois Fleuret

We propose instead to let the model build latent variables

$$Y_r = f_r(S_1, \ldots, S_t, Y_1, \ldots, Y_{r-1}, Z_r; \theta)$$

where Z_r is sampled from a random generator. This can be interpreted as making decisions beside the token choices.

Slide from Francois Fleuret

We propose instead to let the model build latent variables

$$Y_r = f_r(S_1, \ldots, S_t, Y_1, \ldots, Y_{r-1}, Z_r; \theta)$$

where Z_r is sampled from a random generator. This can be interpreted as making decisions beside the token choices.

Sampling from such a trained model is trivial: sample z, and run the AR process as usual to sample $\mathrm{P}_\theta(S \mid Z = z)$.

Training, however, is far more involved.

Slide from Francois Fleuret

The Variational Autoencoder

Slide from Francois Fleuret

Given a training sample s, we want to maximize

$$\mathrm{P}_\theta(S = s) = \mathbb{E}_{z \sim \mathrm{P}(Z)}\Big[\mathrm{P}_\theta(S = s \mid Z = z)\Big]$$

Slide from Francois Fleuret

Given a training sample s, we want to maximize

$$\mathrm{P}_\theta(S = s) = \mathbb{E}_{z \sim \mathrm{P}(Z)}\Big[\mathrm{P}_\theta(S = s \mid Z = z)\Big]$$

Since $\mathrm{P}_\theta(S = s \mid Z = z)$ is a full-fledged model, there is no closed form for $P_\theta(S = s)$, and numerical integration requires "good" Zs.

Slide from Francois Fleuret

Given a training sample s, we want to maximize

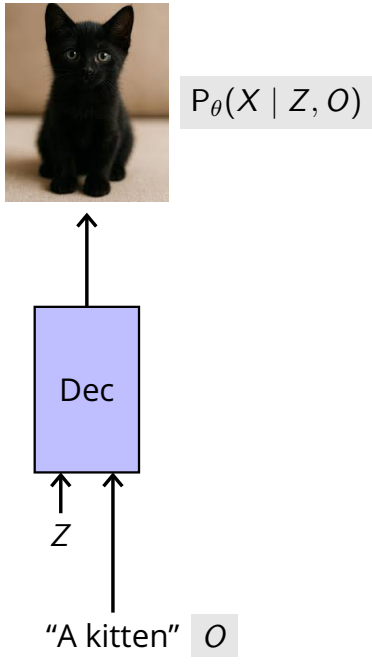
$$\mathrm{P}_\theta(S = s) = \mathbb{E}_{z \sim \mathrm{P}(Z)}\Big[\mathrm{P}_\theta(S = s \mid Z = z)\Big]$$

Since $\mathrm{P}_\theta(S = s \mid Z = z)$ is a full-fledged model, there is no closed form for $P_\theta(S = s)$, and numerical integration requires "good" Zs.

The Variational Autoencoder proposed by Kingma and Welling (2013) provides a formal derivation to train a sampler $q_\theta(Z; S = s)$.

Slide from Francois Fleuret

Inference



Slide from Francois Fleuret

Inference

$P_\theta(X \mid Z, O)$

Dec

Z

"A kitten" O

Training

$P_\theta(X \mid Z, O)$

Dec

?

"A kitten" O

Slide from Francois Fleuret

Inference

$P_\theta(X \mid Z, O)$

Dec

Z

"A kitten" O

Training

$P_\theta(X \mid Z, O)$

$q_\theta(Z; X, O)$

Enc

Dec

X

"A kitten" O

Slide from Francois Fleuret

$$\log \mathrm{P}_\theta(S = s)$$

Slide from Francois Fleuret

$$\log \mathrm{P}_\theta(S = s)$$

$$\geq \log \mathrm{P}_\theta(S = s) - \overbrace{\mathbb{D}_{\mathrm{KL}}\Big(q_\theta(Z; S = s) \Big\| \mathrm{P}_\theta(Z \mid S = s)\Big)}^{\text{How bad the sampler is}}$$

Slide from Francois Fleuret

$$\log \mathrm{P}_\theta(S = s)$$

$$\geq \log \mathrm{P}_\theta(S = s) - \overbrace{\mathbb{D}_{\mathrm{KL}}\Big(q_\theta(Z; S = s) \,\Big\|\, \mathrm{P}_\theta(Z \mid S = s)\Big)}^{\text{How bad the sampler is}}$$

$$= \underbrace{\mathbb{E}_{z \sim q_\theta(Z; S=s)}\left[\log \frac{\mathrm{P}_\theta(S = s, Z = z)}{q_\theta(Z = z; S = s)}\right]}_{\text{“Evidence Lower Bound” (aka ELBO)}}$$

Slide from Francois Fleuret

$$\log \mathrm{P}_\theta(S = s)$$

$$\geq \log \mathrm{P}_\theta(S = s) - \overbrace{\mathbb{D}_{\mathrm{KL}}\Big(q_\theta(Z; S = s) \,\Big\|\, \mathrm{P}_\theta(Z \mid S = s)\Big)}^{\text{How bad the sampler is}}$$

$$= \underbrace{\mathbb{E}_{z \sim q_\theta(Z; S=s)}\left[\log \frac{\mathrm{P}_\theta(S = s, Z = z)}{q_\theta(Z = z; S = s)}\right]}_{\text{“Evidence Lower Bound” (aka ELBO)}}$$

$$= \underbrace{\overbrace{\mathbb{E}_{z \sim q_\theta(Z; S=s)}\Big[\log \mathrm{P}_\theta(S = s \mid Z = z)\Big]}^{-\text{cross-entropy}} - \overbrace{\mathbb{D}_{\mathrm{KL}}\Big(q_\theta(Z; S = s) \,\Big\|\, \mathrm{P}(Z)\Big)}^{\text{How much we cheat}}}_{\text{The quantity we maximize during training}}$$

Slide from Francois Fleuret

The model $q_\theta(Z = z; S = s)$ can be envisioned as an encoder that extracts the necessary information from s.

E.g. "s is a negative review, θ should be improved for $z = -1$".

Slide from Francois Fleuret

The model $q_\theta(Z = z; S = s)$ can be envisioned as an encoder that extracts the necessary information from s.

E.g. "s is a negative review, θ should be improved for $z = -1$".

The term

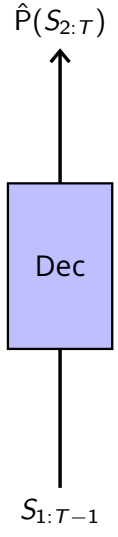
$$\mathbb{D}_{\mathrm{KL}}\Big(q_\theta(Z; S = s) \,\Big\|\, \mathrm{P}(Z)\Big)$$

reflects how much information about s the encoder provides to the decoder, and must be controlled carefully during training.

Slide from Francois Fleuret

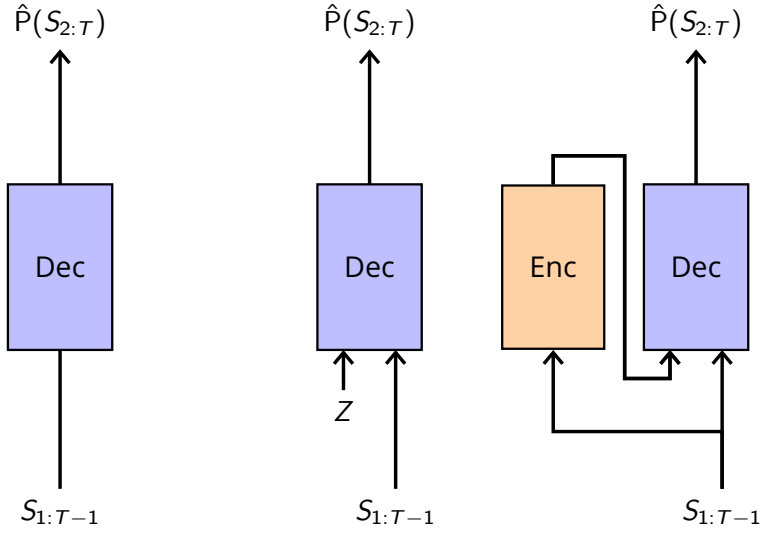
The Free Transformer

Slide from Francois Fleuret



Decoder Transformer

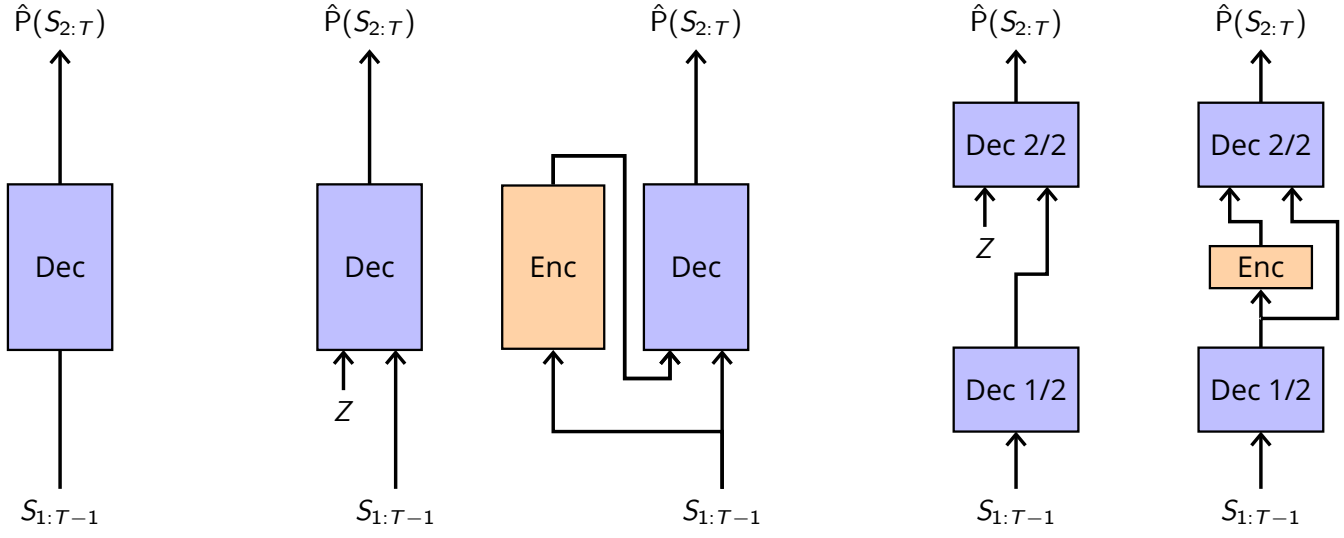
Slide from Francois Fleuret



Decoder Transformer

VAE + Transformer

Slide from Francois Fleuret



Decoder Transformer VAE + Transformer Free Transformer

Slide from Francois Fleuret

We use for Z a one-hot vector of dimension 2^{16}, comparable to the vocabulary size of 2^{17}

Slide from Francois Fleuret

We use for Z a one-hot vector of dimension 2^{16}, comparable to the vocabulary size of 2^{17}

Given $L \in \mathrm{R}^H$, the "Binary Mapper" interprets them as bits logits and

+ Computes the distribution p over $\{0, \ldots, 2^H - 1\}$.
+ Samples $K \sim p$.
+ Outputs $Y_d = \delta_{d=K} + p_d - \mathrm{detach}(p_d)$.

Slide from Francois Fleuret

We use for Z a one-hot vector of dimension 2^{16}, comparable to the vocabulary size of 2^{17}

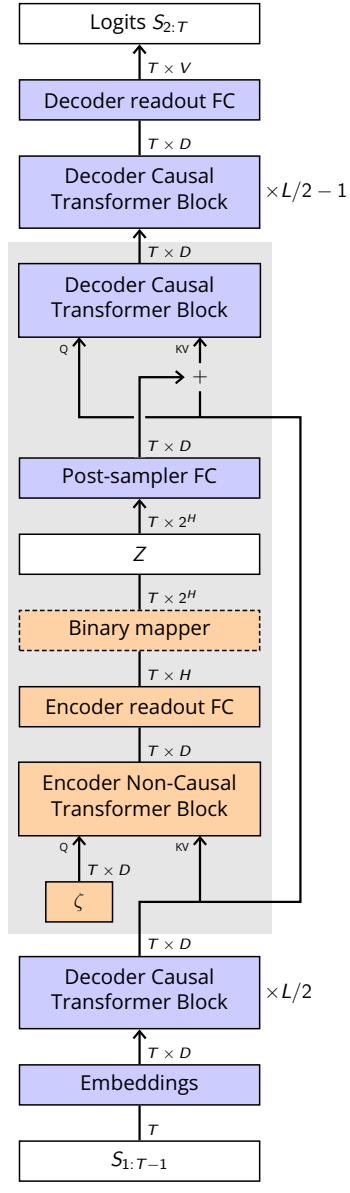
Given $L \in \mathrm{R}^H$, the "Binary Mapper" interprets them as bits logits and

+ Computes the distribution p over $\{0, \ldots, 2^H - 1\}$.
+ Samples $K \sim p$.
+ Outputs $Y_d = \delta_{d=K} + p_d - \mathrm{detach}(p_d)$.

The vector L is:

+ Constant zero for $P(Z)$.
+ The output of the encoder for $q_\theta(Z; S = s)$.

Slide from Francois Fleuret



Slide from Francois Fleuret

The KL divergence can be expressed per token

$$\mathbb{D}_{\mathrm{KL}}\Big(q(Z_t; S_1, \ldots, S_T) \,\Big\|\, \mathrm{P}(Z_t)\Big) = H \log 2 + \sum_{z=1}^{2^H} q(Z_t = z; S) \log q(Z_t = z; S).$$

Slide from Francois Fleuret

The KL divergence can be expressed per token

$$\mathbb{D}_{\mathrm{KL}}\Big(q(Z_t; S_1, \ldots, S_T) \,\Big\|\, \mathrm{P}(Z_t)\Big) = H \log 2 + \sum_{z=1}^{2^H} q(Z_t = z; S) \log q(Z_t = z; S).$$

We use the free-bits method (Kingma et al., 2016), also per token

$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{T} \sum_{t=1}^{T} \max\Big(0,\, \mathbb{D}_{\mathrm{KL}}\Big(q(Z_t; S_1, \ldots, S_T) \,\Big\|\, \mathrm{P}(Z_t)\Big) - \kappa\Big).$$

Slide from Francois Fleuret

Outline

Summary of Future Challenges

The structure of hidden representations will continue to be central to effective deep learning architectures for NLP and AI.

- Recent progress is rooted in unsupervised learning of linguistically appropriate structured representations.
  - Can we learn nodes of the structure (entities)?
  - Can we learn levels of representation?
- Can LLMs learn representations which capture fundamental properties of language?
  - Systematicity, yes.
  - Compositionality?
  - Entailment?
  - Muti-step reasoning?
- What architectural changes are needed to improve the inductive bias of LLMs?
  - Causal random choices within the latent representations?
  - ...