Structural Transfer Learning: Exploring Neural Models and Language Structure Through Understanding Transfer

Isabel Papadimitriou
Transfer Learning

Untrained Model

Training (self-supervised)

Data 1 Model

Data 1

Data 2

Data 2 Model

Similarities

Usually widely available

More specialized

Transfer Learning
Transfer learning has practical applications

But also an **analysis methodology** for understanding data and learning

- Power machine learning models let us explore questions about language in new ways
Shared structures between modalities

Data 1
- Random numbers
- or
- MIDI Music

Data 2
- English wikipedia (~100 mil token subset)

Perplexity on English (lower is better)

<table>
<thead>
<tr>
<th>Pretrained Model</th>
<th>Random</th>
<th>MIDI Music</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>120</td>
<td>40</td>
</tr>
</tbody>
</table>
**Structural Transfer**: a testbed for linguistic structure hypotheses

- Untrained Model
- Structure 1 Model
- English Model

1. **Untrained Model**
2. **Training (self-supervised)**
3. **Linguistic Hypothesis 1**
4. **English Wikipedia**
Structural Transfer: a testbed for linguistic structure hypotheses

- Untrained Model
  - Training (self-supervised)
- Structure 2 Model
- English Model

+ Linguistic Hypothesis 2
+ English Wikipedia
Structural Transfer: a testbed for linguistic structure hypotheses

Untrained Model + Structure 3 Model → English Model

Training (self-supervised)

Linguistic Hypothesis 3 + English Wikipedia

Hypothesis generation about language and language learning
Using structural transfer learning to explore the role of structure in language and language learning
Transfer learning in NLP

- Recent NLP: pretrain so much, that the task can be described in language. **Prompting**

Transfer learning now

- Looking beyond the **dominant languages** where we can do things like prompting
- And for understanding **structure**
Structure and language

- Structure is characteristic of human language
- Most obviously in syntax
- But also beyond syntax
  - Meaning, discourse, reference, information structure
- What structural biases are sufficient for language learning?
- (beyond this talk) Role of communication and language use in creating structure
Three hypotheses about language

1) Recursion

Constituents “Clumping”

The cat sat on the mat

I think that the cat sat on the mat

You always accuse me that I think that the cat sat on the mat
Three hypotheses about language

1) Recursion

Nesting    Context-free
Three hypotheses about language

2) Crossing links and dependencies

Linking in meaning and reference

“I voted for him even though I am negatively affected by his redistribution policies” he said
Three hypotheses about language

2) Crossing links and dependencies

And syntactic structures

... mer  d’chind  em Hans  es huus  lönd  hälle  aastriiche
... we  the children  Hans  the house  let  help  paint

[Schieber 1985]
Three hypotheses about language

3) Zipfian vocabulary distribution
Outline

- What structural biases are useful for human language learners?
  - Disentangling the effects of recursive and linking structures

- How does vocabulary distribution transfer as a structural bias?
  - The structural effect of vocabulary
Outline

● **What structural biases are useful for human language learners?**
  ○ Disentangling the effects of *recursive* and *linking* structures

● **How does vocabulary distribution transfer as a structural bias?**
  ○ The structural effect of vocabulary
Exploring inductive bias

? + Pretty limited linguistic exposure → Language learning

Use transfer learning to test different **structural inductive learning biases**
Transfer learning methodology

Untrained GPT-2 model + Symbolic Language → Structurally-biased learner

Vocabulary

How fast/well can it learn language?
Symbolic pretraining languages

Nesting Parentheses
Nested Parentheses Primitive

- Well-nested, matching pairs
- Constituents
Symbolic pretraining languages

Nesting Parentheses  
Crossing Dependencies
Crossing Dependencies

- Tokens have to match, but not nest
- Where does the structure come from?
  - Dependency length distribution: sample from empirical distances of nesting parentheses
Symbolic pretraining languages

Controls:

- Nesting Parentheses
- Crossing Dependencies
- Random
- Regular repetition
Simple Repetition Primitive

Randomly sample $k$ words, then repeat them, then randomly sample $k$ words…

(Example is for $k=5$, we do $k=10$ in experiments)
Symbolic pretraining languages

Nesting Parentheses

Crossing Dependencies

Random

Regular repetition

Controls:
Nesting structure helps language learning

Baseline - no structure

Baseline - English structure
Multilingual case – Japanese and Basque

Perplexity on Japanese
( lower is better)

Pretrained Model

Random  
Nesting Paren
Control: Pretrained GPT-2

Perplexity on Basque
( lower is better)

Pretrained Model

Random  
Nesting Paren
Control: Pretrained GPT-2
Symbolic pretraining languages

Controls:

- Nesting Parentheses
- Crossing Dependencies
- Random
- Regular repetition
Question: does **nesting** really help? Or would any structure help?
Not just any structure has this effect

Complexity of nesting is necessary

Simple structure still significant
Again, a multilingual effect
Symbolic pretraining languages

Controls:

- Nesting Parentheses
- Crossing Dependencies
- Random
- Regular repetition
Crossing links, without nesting, provide a better inductive bias.
This is also true across languages
The kinds of structure that make language are multifaceted

- Structural transfer lets us explore hypotheses about structure in language
- Language as a learnable system, independent of linguistic theory
Mixing nesting and crossing parentheses

- A language that is mostly nesting, with 1%, or 10% of parentheses not following the structure.
Slightly breaking constituent structure makes better language learners
Also a multilingual effect
Structural inductive bias through transfer learning

- Complex structural relationships are important in language
- Multiple crossing dependencies a good starting point for language learning
- Computational models as **hypothesis generators**: testing linguistic structure in theory-free ways
Outline

● What *structural biases* are useful for human language learners?
  ○ Disentangling the effects of *recursive* and *linking* structures

● How does vocabulary distribution transfer as a structural bias?
  ○ The structural effect of vocabulary
Outline

● What structural biases are useful for human language learners?
  ○ Disentangling the effects of recursive and linking structures

● How does vocabulary distribution transfer as a structural bias?
  ○ The structural effect of vocabulary
The lexicon in linguistics

- A good amount of structure is in the vocabulary:
  - Vocabulary distribution
  - Structure in meaning
  - and also in grammar
    - Properties like transitive verb
    - Constructions, like “Let alone”
    - …
We throw out the vocabulary between pretraining and fine-tuning

Pretraining vocabulary

Vocabulary indices 0-499

What the model sees

Fine-tuning vocabulary

cat, dog … book … ameliorate …

Vocabulary indices 0-50K
Structural Transfer: a testbed for linguistic structure hypotheses

Untrained Model + Vocabulary Distribution 1 → Structure 1 Model
Training (self-supervised) + English Wikipedia → English Model
Does a Zipfian vocabulary distribution in pretraining have a **structural** effect?

- Even though we discard vocabulary information
Yes, Zipfian information is transferred
... but does not necessarily combine with structure
The role of vocabulary in transfer learning is an interesting problem.

Vocabulary embedding matrix

Lexical representations

“The cat sat on the mat”

$|V| \approx 10s$ of K

Really big transformer model
The role of vocabulary in transfer learning is an interesting problem

- A practical problem: without enough data, it’s hard to see a word often enough to learn a good vector
- A puzzle: how is structural information separated between vocabulary matrix and model weights?
  - Vocabulary information like distribution can have structural effects
Transfer learning, language, and structure

- Transfer learning is a test bed for understanding structure in language learning

- Computational models of cognitive processes can’t prove anything – but they serve as interesting hypotheses generators

- It’s an exciting time: machine learning opens up new avenues for exploring questions in language

Thanks!