Unsupervised Distillation Of Syntactic Information From Contextualized Word Representations

Shauli Ravfogel*, **Yanai Elazar***, Jacob Goldberger and Yoav Goldberg

Presented at: BlackBox NLP, EMNLP 2020

CMU, 29th November, 2020

Language Is Complex

- Human language is a complex system, involving an intricate play between structure and meaning

```
“One morning, I shot an elephant in my pajamas.

How he got into my pajamas I'll never know.”
```

Language Is Complex

- Consider the following sentences:

Language Is Complex

- Consider the following sentences:

Green ideas are colorless

Language Is Complex

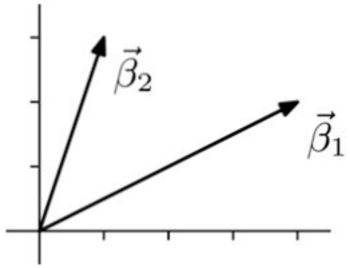
- Consider the following sentences:

Green ideas are colorless

Neural networks are interesting

Language Is Complex

- Consider the following sentences:
- Although the sentences convey a different meaning

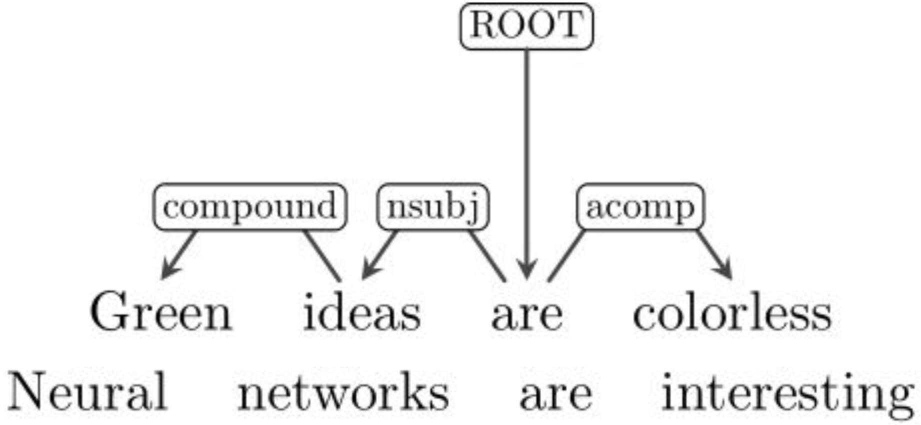


Green ideas are colorless

Neural networks are interesting

Language Is Complex

- Consider the following sentences:
- Although the sentences convey a different meaning
- Their structure is alike



Language Is Complex

Do LMs capture this complexity?

LMs capture language!

- Impressive performance on syntactic and semantic tasks

Gulordava et al., 2018; Tenney et al, 2019; Yang et al, 2019

LMs capture language!

- Impressive performance on syntactic and semantic tasks
- Encoding syntax with no explicit supervision

Goldberg, 2019; Liu et al, 2019; Clark et al, 2019; Hewitt and Manning, 2019

LMs capture language!

- Impressive performance on syntactic and semantic tasks
- Encoding syntax with no explicit supervision
- Can we separate semantics from syntax?

This work!

Disentanglement

- Disentanglement is the differentiation between different types of information encoded in a representation.

Disentanglement

- Disentanglement is the differentiation between different types of information encoded in a representation.

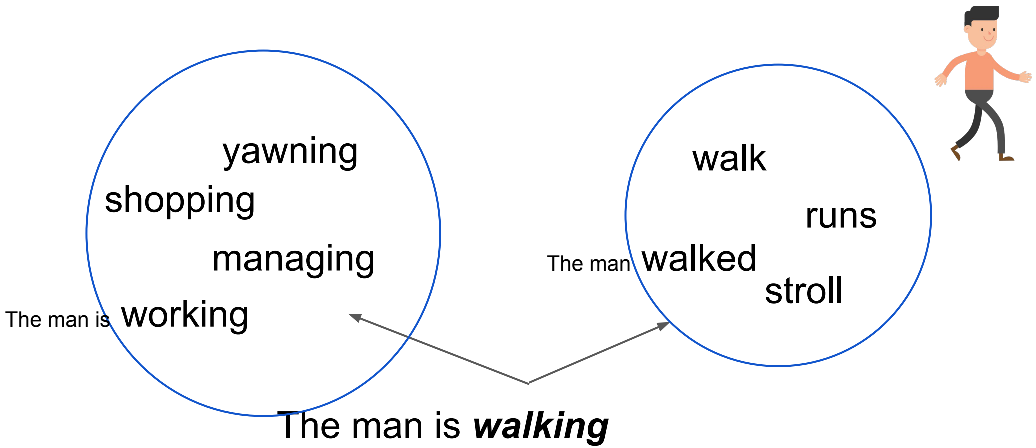
The man walked

The man is working

The man is ***walking***

Disentanglement

- Disentanglement is the differentiation between different types of information encoded in a representation.



Disentanglement

- Disentanglement is the differentiation between different types of information encoded in a representation.
- Disentanglement between syntactic and semantic representations is often a desired property:

Disentanglement

- Disentanglement is the differentiation between different types of information encoded in a representation.
- Disentanglement between syntactic and semantic representations is often a desired property:
    - Can we understand a model behavior & mistakes

Disentanglement

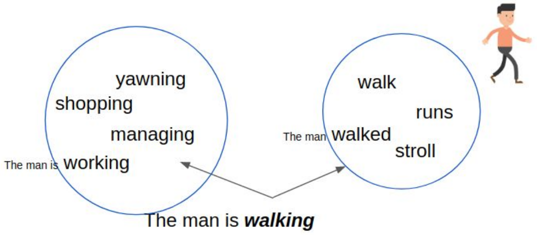
- Disentanglement is the differentiation between different types of information encoded in a representation.
- Disentanglement between syntactic and semantic representations is often a desired property:
  - Can we understand a model behavior & mistakes
  - We often want to achieve *invariance* to one kind of information, while keeping the other:
    - E.g. saying the same “content” in a different “style”

Why separate syntax from semantics?

- Can **discard** the syntactic part, leading to representations which are invariant to syntactic differences
- Can **keep** only the syntactic part, allowing to more cleanly investigates the way LMs handel structure in language

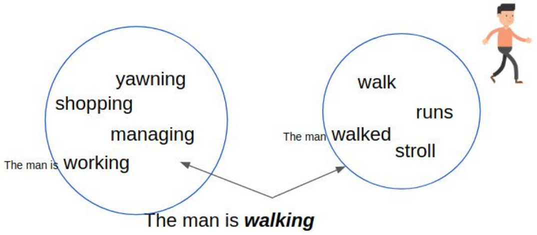
Disentanglement - Objective

- In this work, we focus on disentanglement in LMs
- Given a LM, we want to distill from its representations only those part that capture structure



Disentanglement - Objective

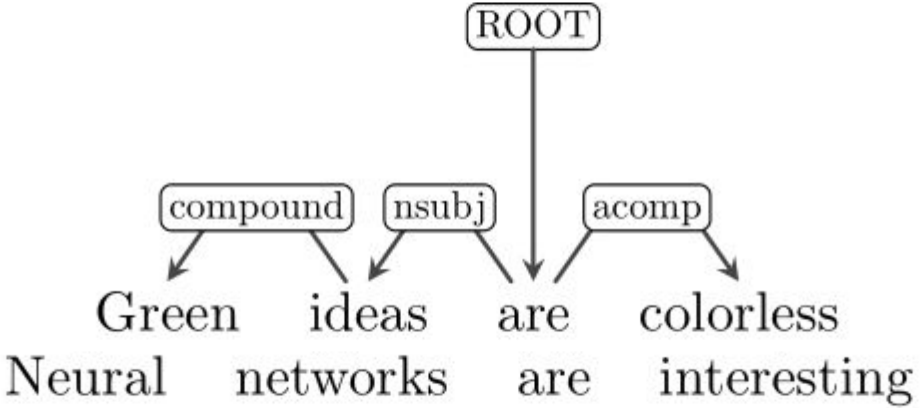
- In this work, we focus on disentanglement in LMs
- Given a LM, we want to distill from its representations only those part that capture structure
- In an unsupervised fashion:
  - We don't assume a specific syntactic scheme



Why unsupervised?

- The syntactic representations of the model don’t necessarily align with any specific scheme
- Probing work has demonstrated limitations of the supervised setting as a way to evaluate the model’s syntactic abilities.

Disentanglement - Objective



- Learn a transformation *f*, where:
  - $f(v_{Neural}) \simeq f(v_{Green})$
  - $f(v_{networks}) \simeq f(v_{ideas})$
  - …

Approach

- Given a dataset of parallel sentence with similar structure

High	school	is	boring
Green	ideas	are	colorless
Neural	networks	are	interesting

Parallel sentences

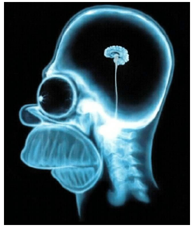
Approach

- Given a dataset of parallel sentence with similar structure

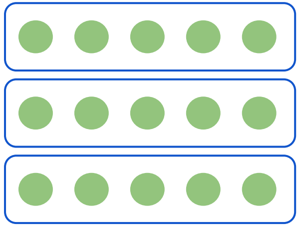
High school is boring
Green ideas are colorless
Neural networks are interesting

Parallel sentences

Language Model



Our Model



Representation

<u>Should be similar</u>

Approach

- Given a dataset of parallel sentence with similar structure

High	school	is	boring
Green	ideas	are	colorless
Neural	networks	are	interesting

Parallel sentences

But how can we get these sentences???
(remember, no supervision)

Approach

- Our solution: use an LM to create alternatives

Green ideas are colorless

Green [mask] are colorless

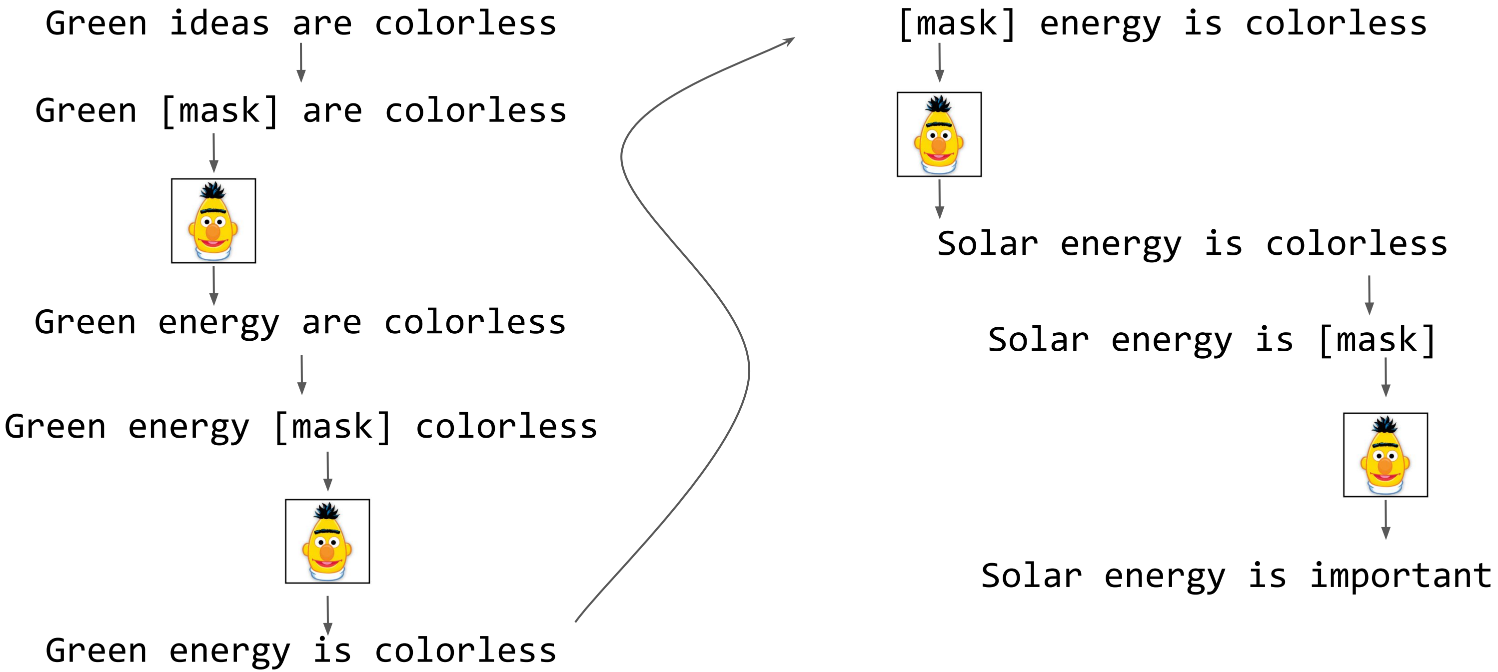
Green energy are colorless

Green energy [mask] colorless

Green energy is colorless

Approach

- Our solution: use an LM to create alternatives



Parallel Syntactic Sentences

- We sample 150K sentence from wikipedia for a starting seed
- and employ our process to generate 5 parallel sentences for each original sentence

When a train ticket is purchased, a contract is established
When a travel document is acquired, a settlement is declared
When a winning vehicle is obtained, a competition is introduced
When a winning bid is announced, a winner is created

Learning a Syntactic Representation

- Using the parallel syntactic corpus

High	school	is	boring
Green	ideas	are	colorless
Neural	networks	are	interesting

Learning a Syntactic Representation

- Using the parallel syntactic corpus
- We can learn a metric *f* such that:

High	school	is	boring
Green	ideas	are	colorless
Neural	networks	are	interesting

Learning a Syntactic Representation

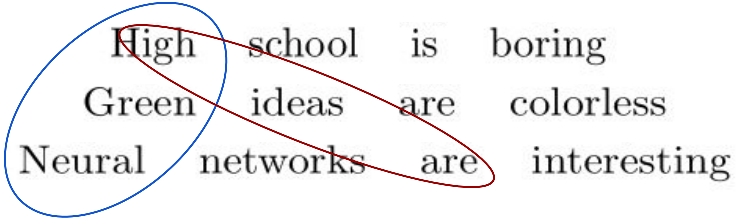
- Using the parallel syntactic corpus
- We can learn a metric *f* such that:
  - words of the same function are close

High	school	is	boring
Green	ideas	are	colorless
Neural	networks	are	interesting

f('High') ≃ *f*('Green') ≃ *f*('Neural')

Learning a Syntactic Representation

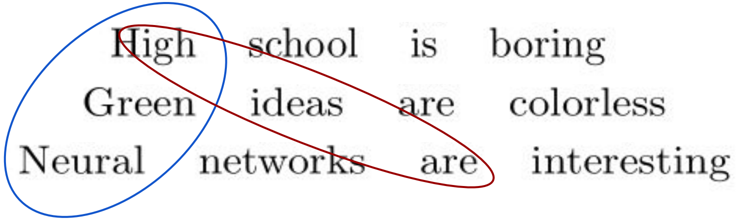
- Using the parallel syntactic corpus
- We can learn a metric *f* such that:
  - words of the same function are close
  - otherwise, they should be distant



f('High') ≃ *f*('Green') ≃ *f*('Neural')
f('High') ≄ *f*('ideas) ≄ *f*(are')

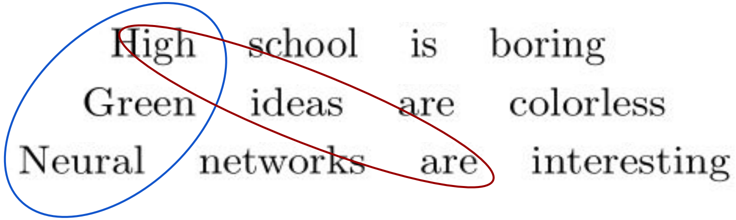
Learning a Syntactic Representation

- In practice, out of the parallel sentences,
  - we use words of same indices as positive examples
  - and some words as negative examples



Learning a Syntactic Representation

- In practice, out of the parallel sentences,
  - we use words of same indices as positive examples
  - and some words as negative examples
- The transformation *f* is a simple function: a matrix mapping to dimensionality of 75.



Learning a Syntactic Representation

- The challenge:
  - There are many negative examples
  - Many would be easy to separate
  - Hard to learn a meaningful representation
- The solution:
  - Use a Triplet-loss objective to mine the “hard examples”

Triplet Loss

- Given a batch with parallel sentences

Group1	Group2
Green ideas are colorless	Who proposed this idea ?
Solar energy is important	What helped the helpless man?

Triplet Loss

- Given a batch with parallel sentences
- Choose an “anchor” word $\boldsymbol{V}^{\mathbf{A}}$:

Group1	Group2
`Green ideas are colorless`	`Who proposed this idea ?`
`Solar energy is important`	`What helped the helpless man?`

Triplet Loss

- Given a batch with parallel sentences
- Choose an “anchor” word V^A:
- Sample a word from the same group, in the same index to be a positive example V^P

Group1	Group2
Green $\boxed{\text{ideas}}$ are colorless	Who proposed this idea ?
Solar $\boxed{\text{energy}}$ is important	What helped the helpless man?

Triplet Loss

- Given a batch with parallel sentences
- Choose an "anchor" word V^{A}:
- Sample a word from the same group, in the same index to be a positive example V^{P}
- Choose the closest word (after the transformation) from the batch to be the negative example V^{N}

Group1	Group2
Green ideas are colorless	Who proposed this idea ?
Solar energy is important	What helped the helpless man?

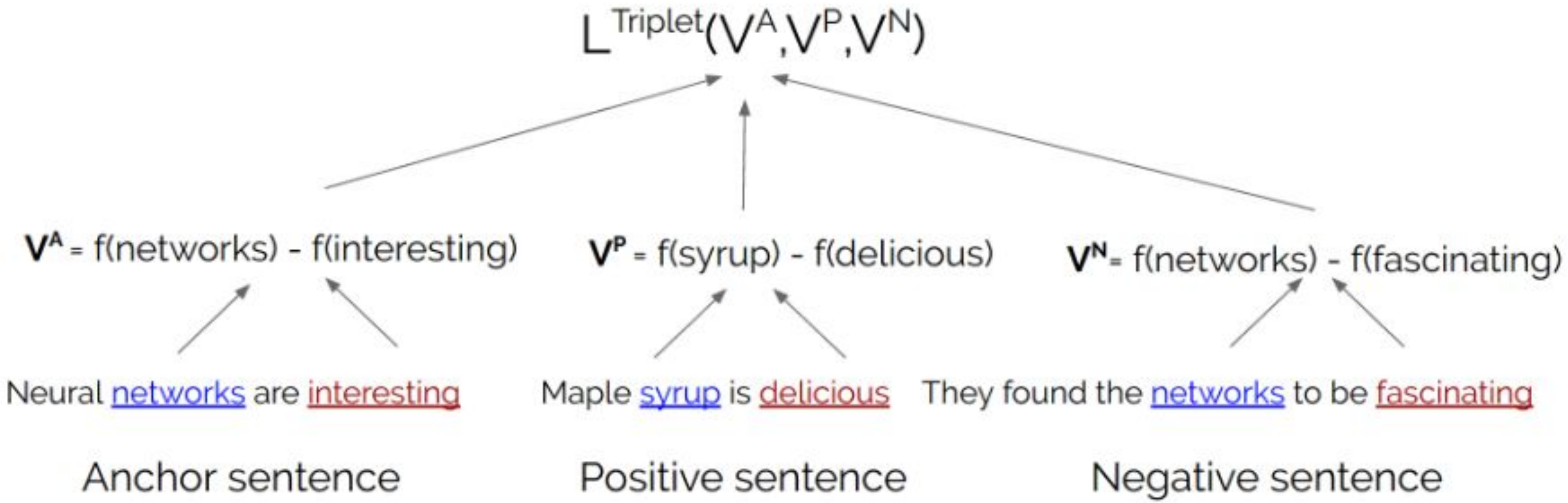
Triplet Loss

- Given a batch with parallel sentences
- Choose an “anchor” word $\boldsymbol{V^A}$:
- Sample a word from the same group, in the same index to be a positive example $\boldsymbol{V^P}$
- Choose the closest word (after the transformation) from the batch to be the negative example $\boldsymbol{V^N}$
- Optimize:

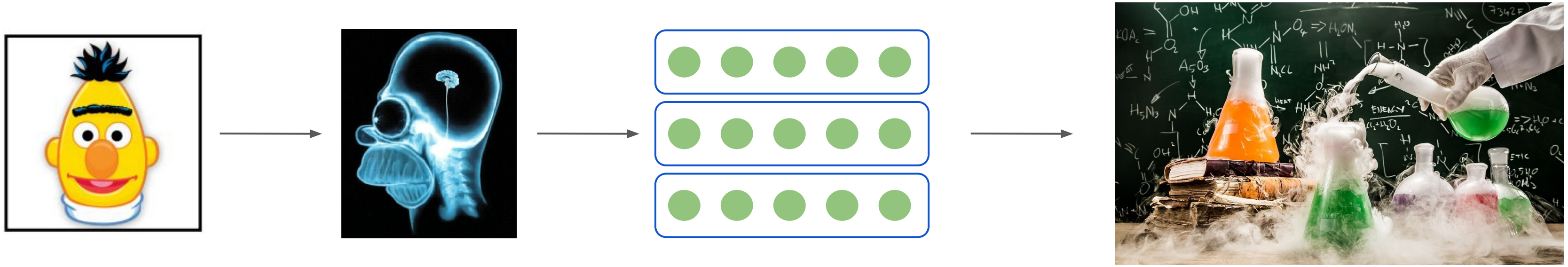
$$L^{triplet}(V^A, V^P, V^N) = \frac{e^{dist(V^A, V^P)}}{e^{dist(V^A, V^P)} + e^{dist(V^A, V^N)}}$$

Metric Learning & Triplet Loss

- We pose the syntax-distillation objective as a metric learning problem.
- We want to learn f that induces a metric under which the representations of structurally-equivalent pairs are close in space.



Experiments and Analysis



Experiments and Analysis

- To evaluate the learned transformation, we check:
  - What was captured in the representations?
  - Are these representations any good?

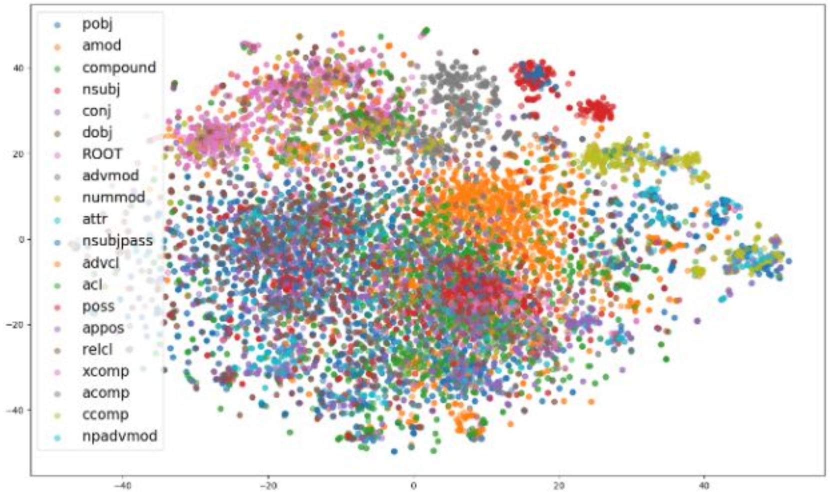
Experiments and Analysis

- We evaluate the learned transformation using:
  - Analysis in the representations space:
    - Are structurally-equivalent words close in space?
    - Does the representation space reflects syntactic relations?
  - Low resource parsing

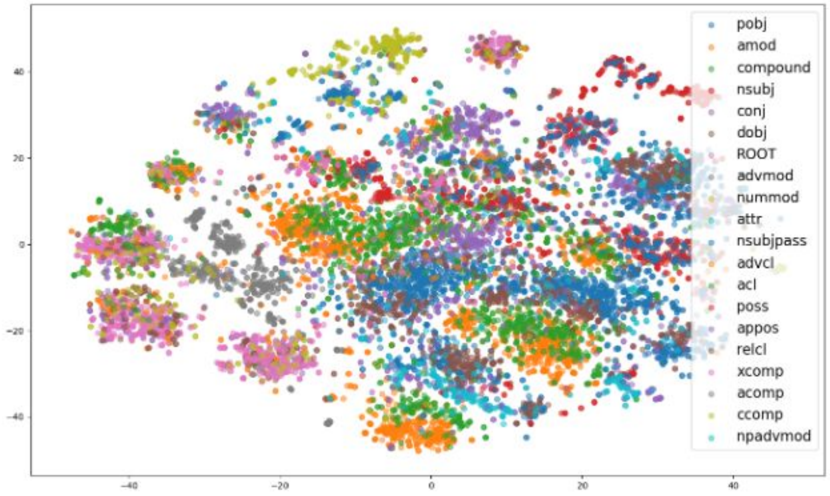
Qualitative Analysis

- We sample words, and look for their nearest neighbors

Elmo representation



Transformed representation



Purity of 80 unsupervised clusters increases from 36.4 to 48.0%

Closest-word query

```
the mint’s director at the time, nicolas peinado, was also an
architect and made the initial plans
```

Closest-word query

the mint’s **director** at the time, nicolas peinado, was also an architect and made the initial plans

Closest-word query

the mint's **director** at the time, nicolas peinado, was also an architect and made the initial plans

Closest-word query

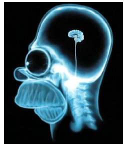
the mint’s **director** at the time, nicolas peinado, was also an architect and made the initial plans

Closest vector

the **director** is angry at crazy loop and glares at him, even trying to get a woman to kick crazy loop out of the show (which goes unsuccessfully).

Closest-word query

`the mint's director at the time, nicolas peinado, was also an architect and made the initial plans`

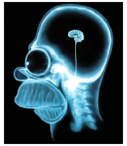


Closest vector

`the director is angry at crazy loop and glares at him, even trying to get a woman to kick crazy loop out of the show (which goes unsuccessfully).`

Closest-word query

the mint’s **director** at the time, nicolas peinado, was also an architect and made the initial plans



Closest vector

Closest vector

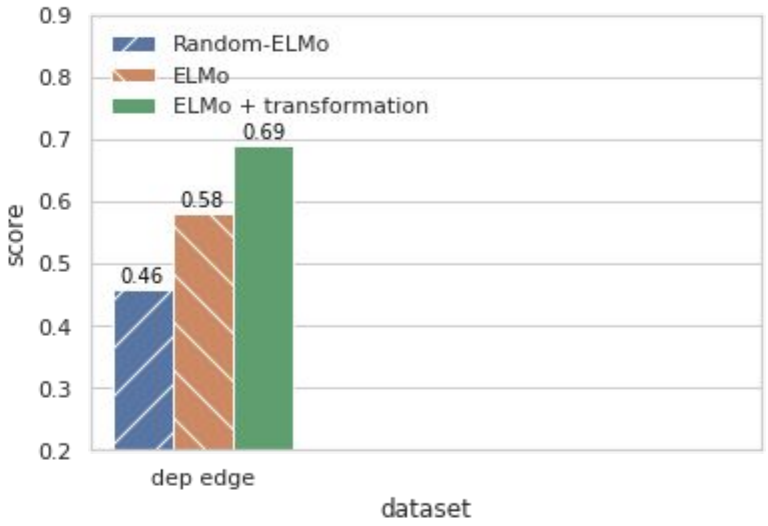
the **director** is angry at crazy loop and glares at him, even trying to get a woman to kick crazy loop out of the show (which goes unsuccessfully).

jetley’s **mother**, kaushaliya rani, was the daughter of high court advocate shivram jhingan.

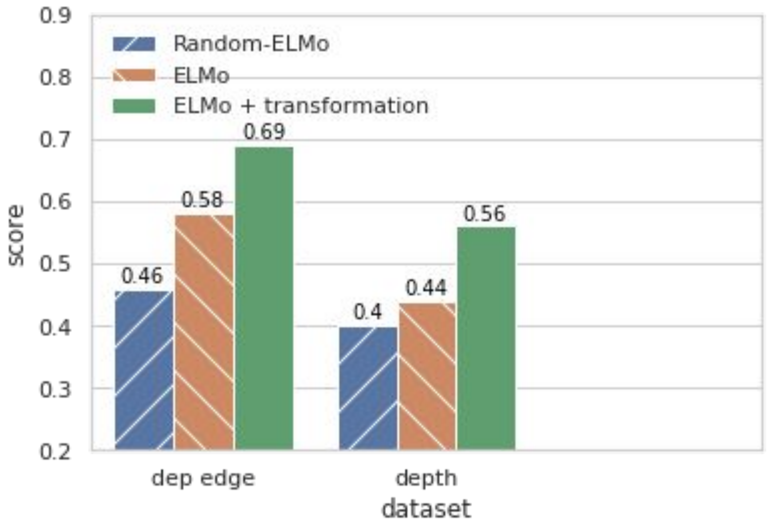
Quantitative results

- Closest words: structural probes:
  - Local structure: dep edge (accuracy match)
  - Depth (correlation)
  - Lexical match (accuracy match)
- Multiple baselines:
  - Random ELMo
  - ELMo

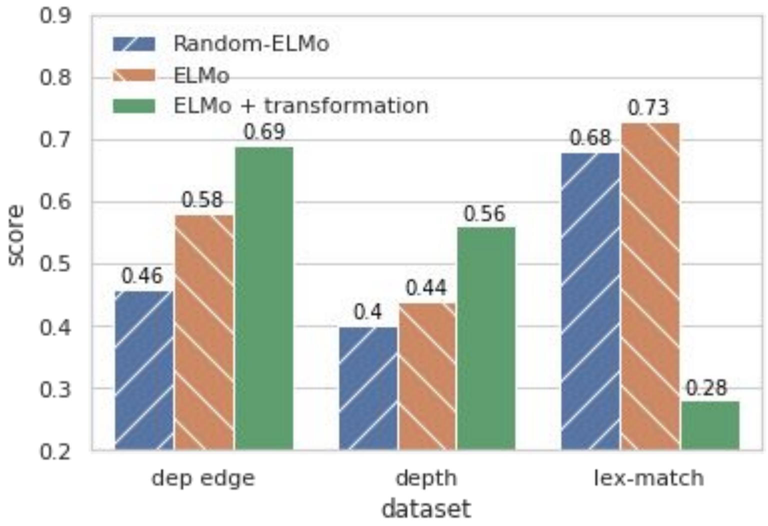
Quantitative results



Quantitative results



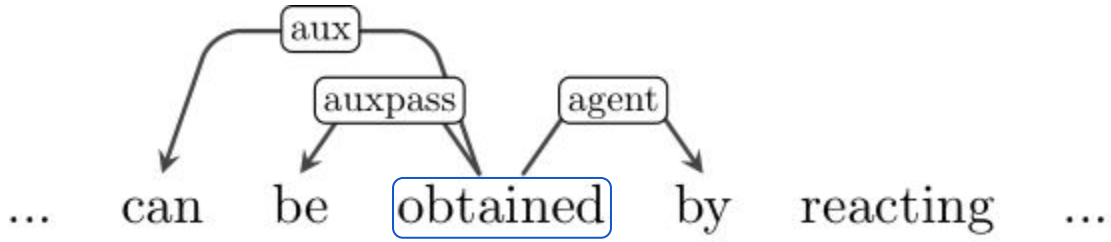
Quantitative results



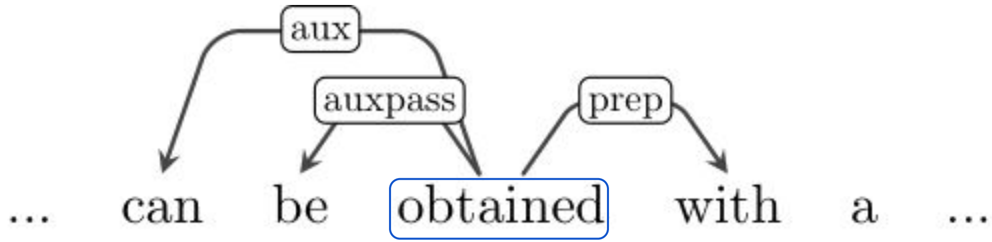
Distilling ELMo For Parsing

- Shift from “traditional” syntactic schemas

Query



Our nearest



Distilling ELMo For Parsing

- Shift from “traditional” syntactic schemas
- How close are these representations to “traditional” schemas?

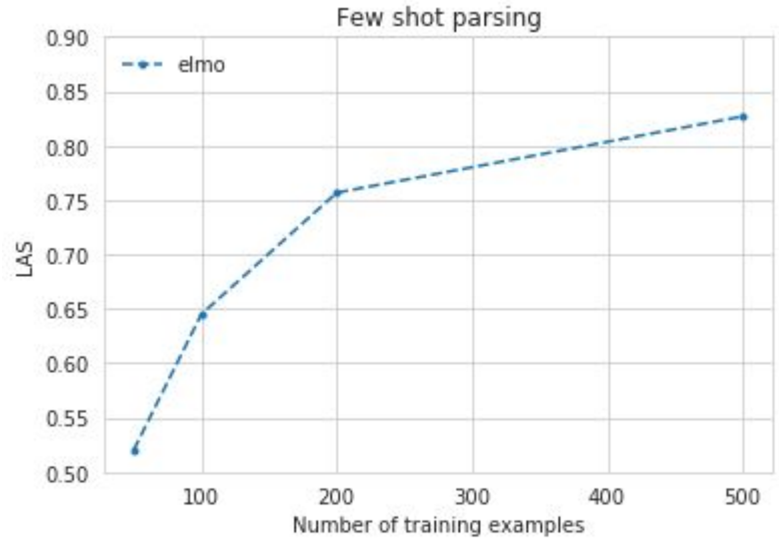
Distilling ELMo For Parsing

- Shift from “traditional” syntactic schemas
- How close are these representations to “traditional” schemas?
- We train a dependency parser over our representations in the low-data regime

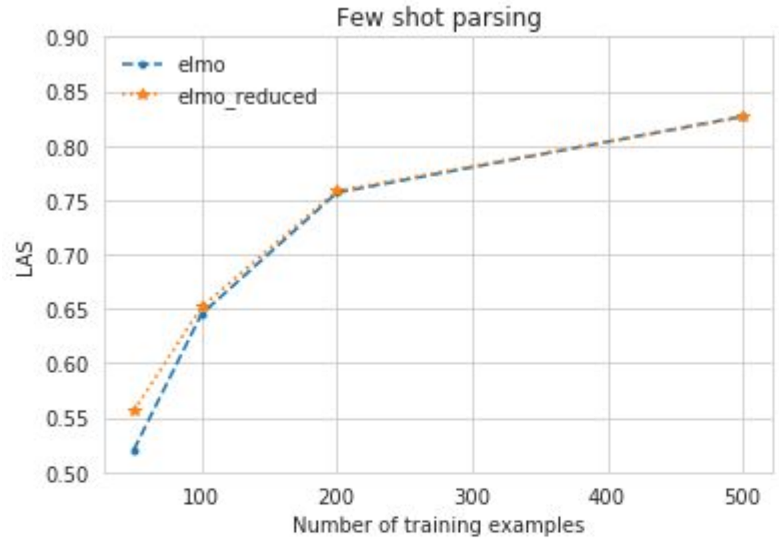
Distilling ELMo For Parsing

- How close are these representations to “traditional” schemas?
- We train a dependency parser over our representations in the low-data regime

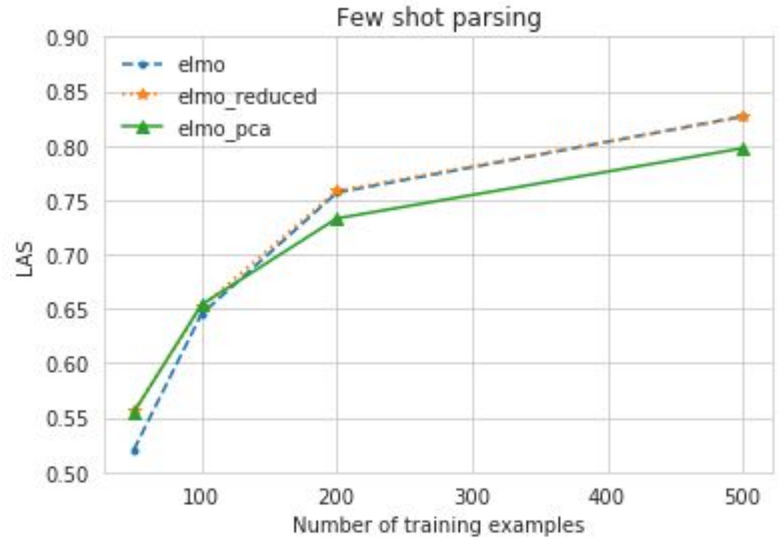
Quantitative results: Parsing



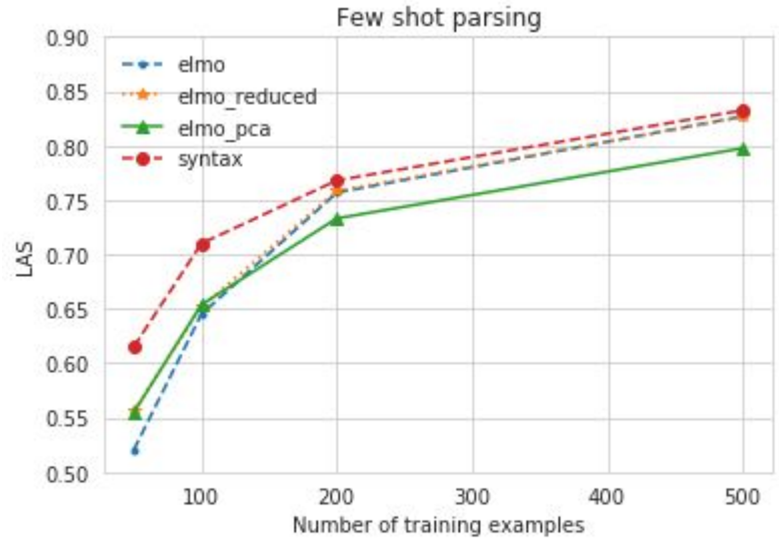
Quantitative results: Parsing



Quantitative results: Parsing



Quantitative results: Parsing



Discussion

- What kind of structure did we learn exactly?
- Can we generate structurally-equivalent sentences which are not of the same length?
  - This requires filling a phrase in the place of a single word.
- Can we get groups of sentences that say the same thing in a different structure?

Conclusions

- We introduce a method for automatic generation of syntactically-equivalent sentences
- We propose an unsupervised approach for extracting structure of language
- We have shown that our representation:
    - Clusters words by structural function
    - Is useful for structural end-tasks

Thanks!
Questions?