Learning Challenges in NLP

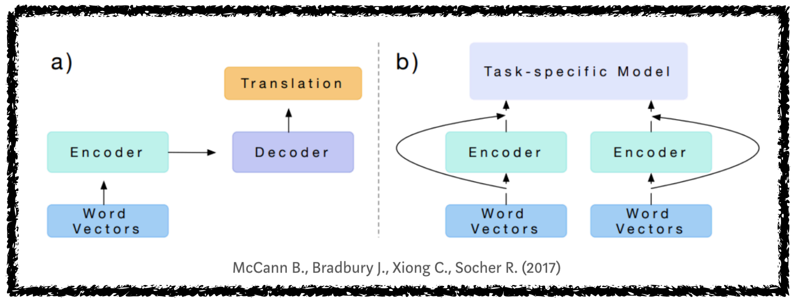
Swabha Swayamdipta

Mar 11, 2019

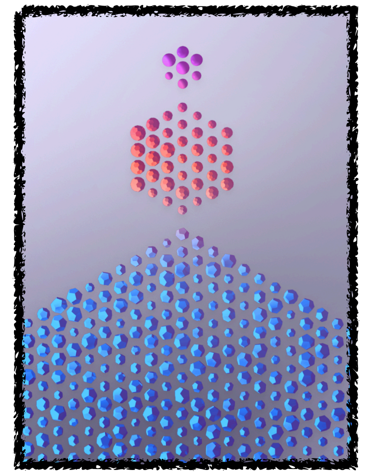
UWNLP

Carnegie Mellon University
Language Technologies Institute

NLP today



[McCann et. al., 2017]



[Radford et. al., 2018]

Large Datasets

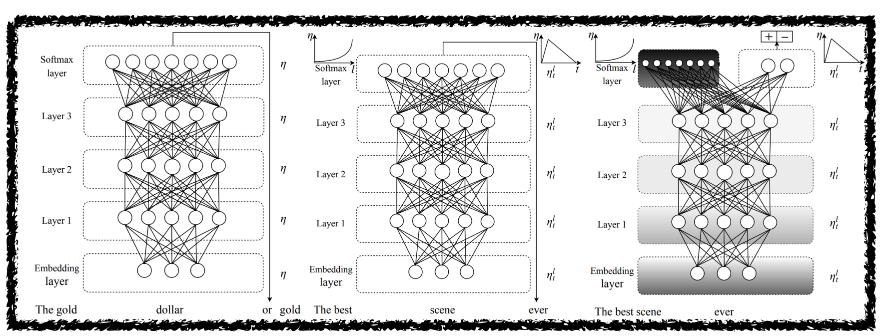
Unsupervised

[Peters et. al., 2018]

Contextualized Representations

Supervised

Downstream Tasks



[Howard & Ruder, 2018]

[Devlin et. al., 2018]

A closer look...

On 31 December 1687 the first organized group of Huguenots set sail from the Netherlands to the Dutch East India Company post at the Cape of Good Hope. The largest portion of the Huguenots to settle in the Cape arrived between 1688 and 1689 in seven ships as part of the organised migration, but quite a few arrived as late as **1700**; thereafter the numbers declined and only small groups arrived at a time.
The number of old Acadian colonists declined after the year **1675**.

The number of new Huguenot colonists declined after what year?

[Jia & Liang, 2017]

BERT [Devlin et. al., 2018]

Learning Challenges

Part I

Can we incorporate some priors about language into deep learning models?

- ☐ Syntactic Scaffolds for Semantic Structures (EMNLP 2018)

Part II

What in our data is causing models to achieve high performance?

- ☐ Annotation Artifacts in Natural Language Inference Data (NAACL 2018)

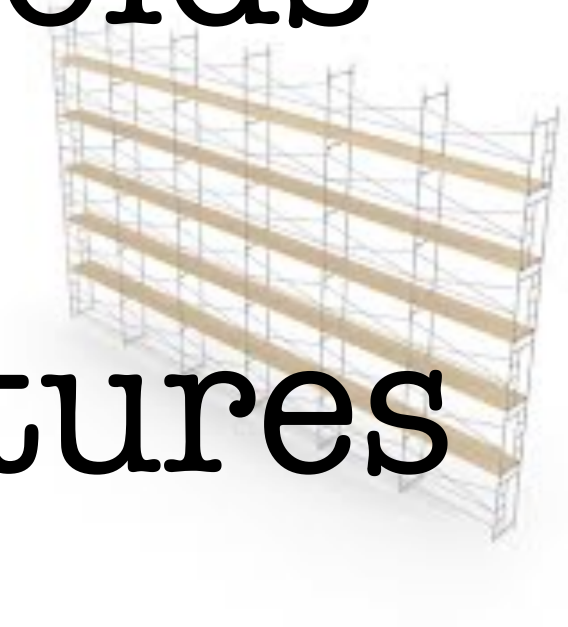
Learning Challenge #1

- Can we incorporate some priors about language?
- One kind of prior - Linguistic Structure
- Can linguistic structure act as an informative prior for models of deep learning?

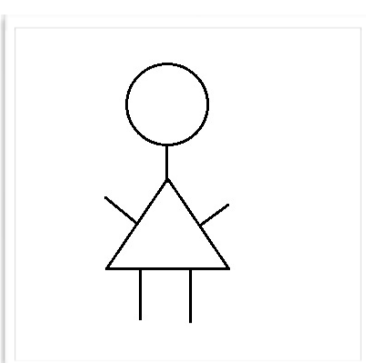
On 31 December 1687 the first organized group of Huguenots set sail from the Netherlands to the Dutch East India Company post at the Cape of Good Hope. The largest portion of the Huguenots to settle in the Cape arrived between 1688 and 1689 in seven ships as part of the organised migration, but quite a few arrived as late as **1700**; thereafter the numbers declined and only small groups arrived at a time.

The number of new Huguenot colonists declined after what year?

Syntactic Scaffolds for Semantic Structures



EMNLP 2018


S.


Sam Thomson


Kenton Lee


Luke Zettlemoyer


Chris Dyer


Noah A. Smith

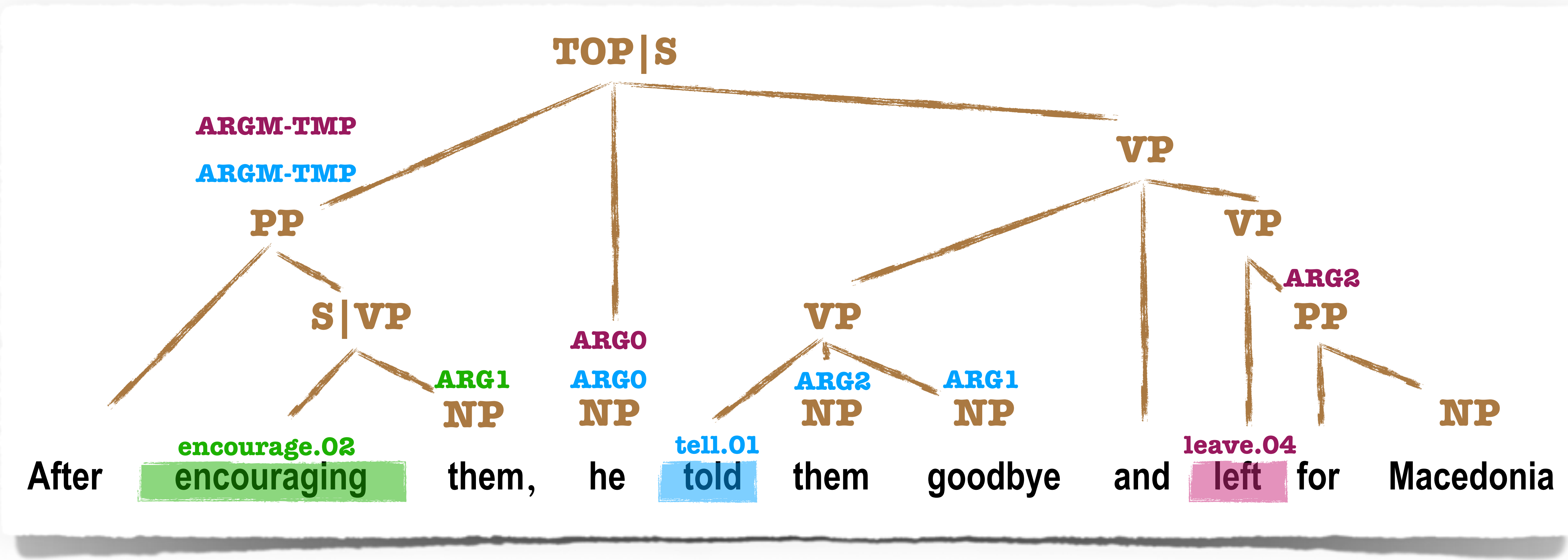
Linguistic Structure: Semantics

- Who did what to whom?
- This talk: **Span**-based, broad-coverage semantic structured prediction.
- Availability of data…

PropBank [Palmer et. al., 05]

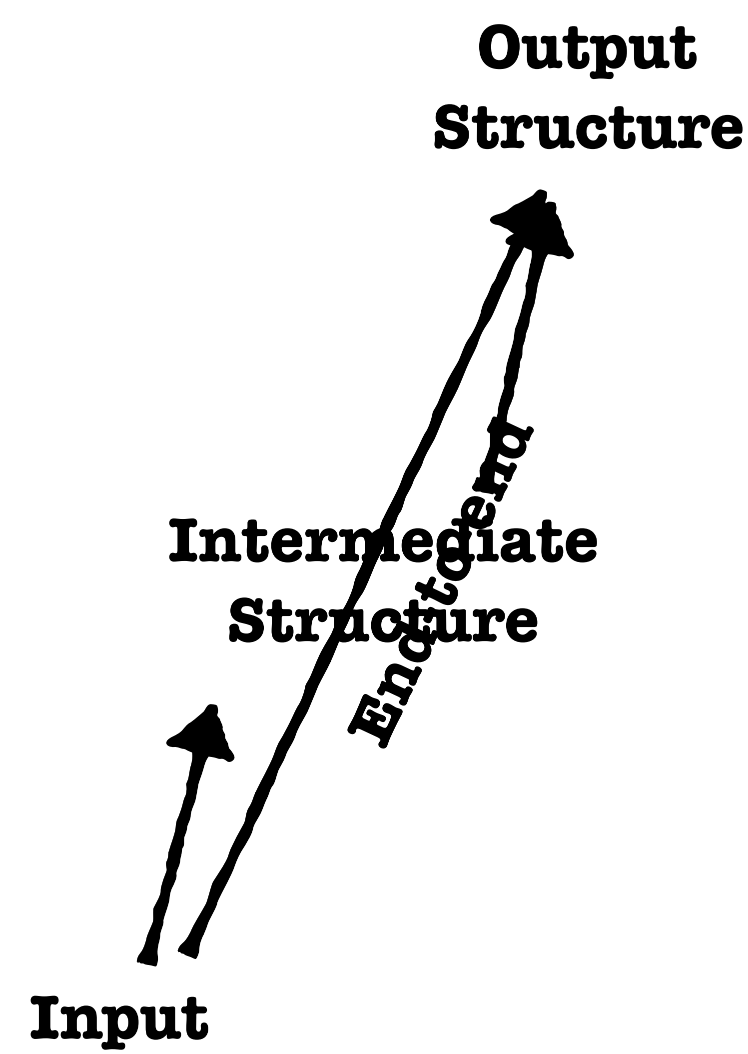
Syntax 💘 Semantics

- Syntax - a foundation for sentence meaning / semantics
- Phrase-based syntax (node → span)
- Key Intuition: Learning from **multiple, complementary** structures results in stronger representations.



Structured prediction with intermediate structures

- Intermediate syntactic structures.
- Traditionally a pipeline, both at train and test time.
    - More structured data
    - Cascading errors
- Forsaken in most end-to-end models, but at a cost [He et. al, 2017].



Training Paradigms

Syntax-free training

End-to-end modeling [He et. al., 2017]

Learning Joint Semantic Parses from Disjoint Data. (NAACL 2018)

Latent variables for syntax [Naradowsky et. al. 2013]

Syntax for training

Multitask Parameter Sharing [Søgaard et. al., 2016]

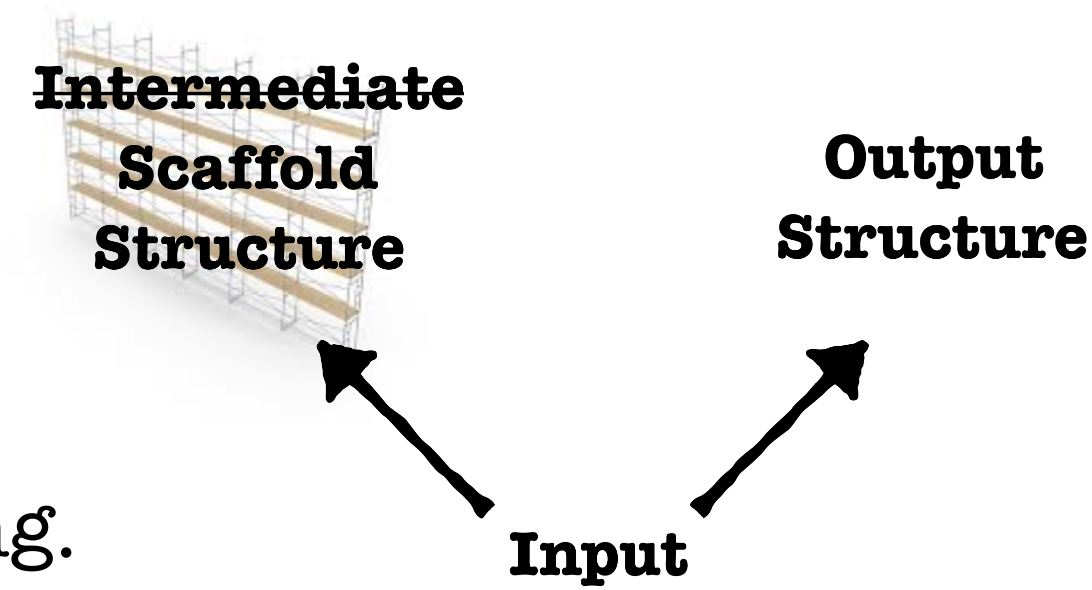
Joint Modeling [Hendersen et. al., 2013]

Syntactic Pipelines [Gildea & Jurafsky, 2002]

Difficulty of designing and training

FREDA [Daumé III, 2009]

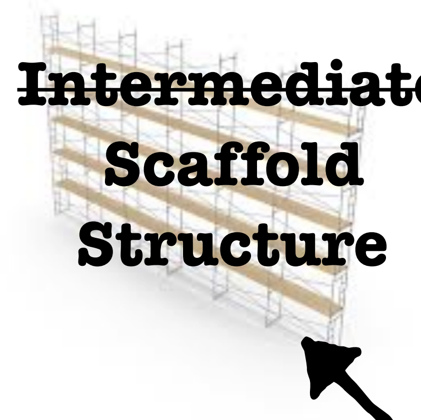
Syntactic Scaffolds



- Multitask setting.
- Shallow intermediate structure prediction.
- Learn (soft) syntax-aware representations, avoid cascaded errors.
- Not required during test.

Task Overview

(Phrase-based syntax)

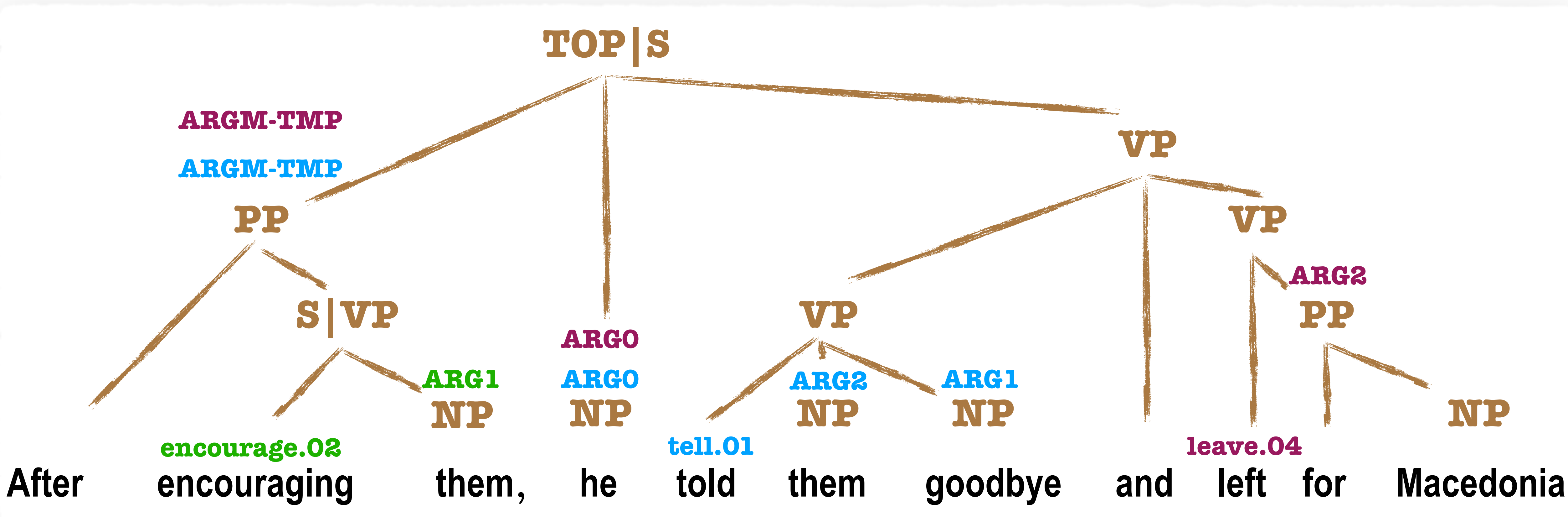


Intermediate Scaffold Structure

Output Structure

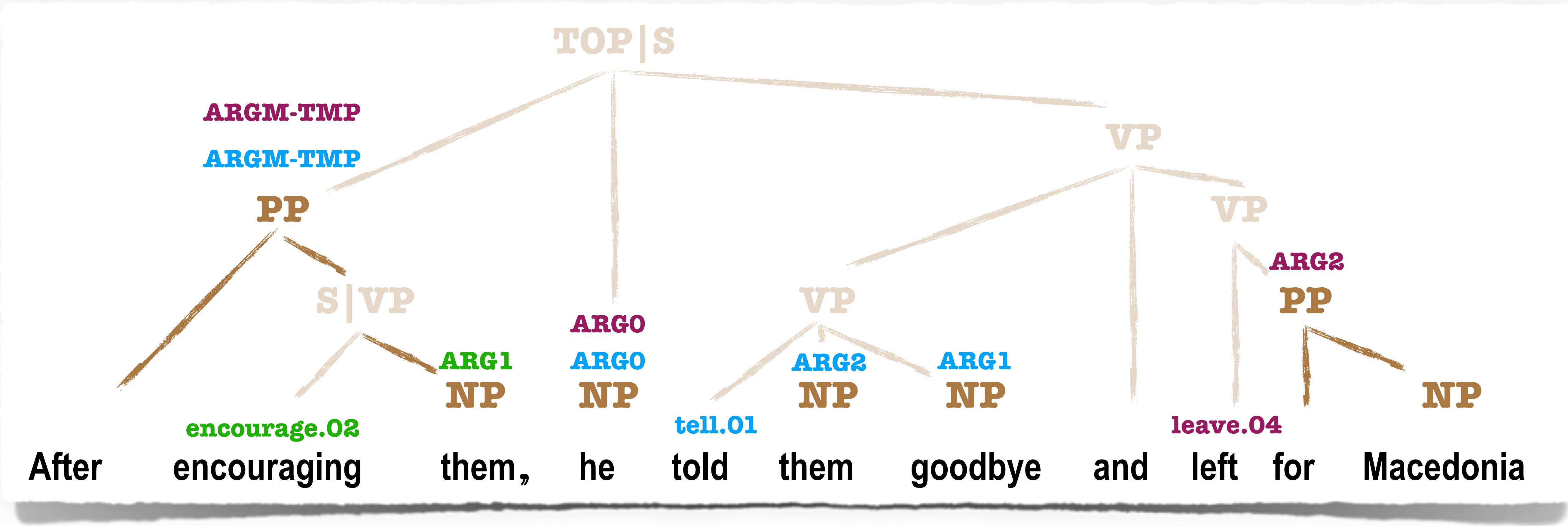
Input
(Sentences or Documents)

- ☑ Frame-Semantic Role Labeling
- ☑ PropBank Semantic Role Labeling
- ☑ Coreference Resolution



How to build a syntactic scaffold?

- **Desired** parts of syntactic tree:



- Span-level classification $$\mathcal{L}_2(\mathbf{x}, \mathbf{z}) = -\sum_{1 \leqslant i \leqslant j \leqslant n} \log p(z_{i:j} \mid \mathbf{x}_{i:j}) \,.$$

Training with syntactic scaffolds

x = Input
y = Output Structure
z = Scaffold Structure

Mixing Ratio

$$\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_1} \mathcal{L}_1(\mathbf{x},\mathbf{y};\theta,\phi) + \delta \sum_{(\mathbf{x},\mathbf{z})\in\mathcal{D}_2} \mathcal{L}_2(\mathbf{x},\mathbf{z};\theta,\psi)$$

Primary Task Objective

Scaffold Task Objective

Primary Dataset

Scaffold Dataset

Shared parameters for input representations

The primary objective

Same structures must be scored in both the primary and the scaffold task.

- Span-based classification, with aggressive pruning [Lee et. al., 2017].
- Semi-Markov Conditional Random Fields [Sarawagi et. al. 2004].

Semi-Markov CRFs

After encouraging them he told them goodbye and left for Macedonia

ARGM-TMP ARG0 leave.04 ARG2

- Globally normalized model for segmentations (**s**) of a sentence (**x**).

$$p(\mathbf{s} \mid \mathbf{x})$$

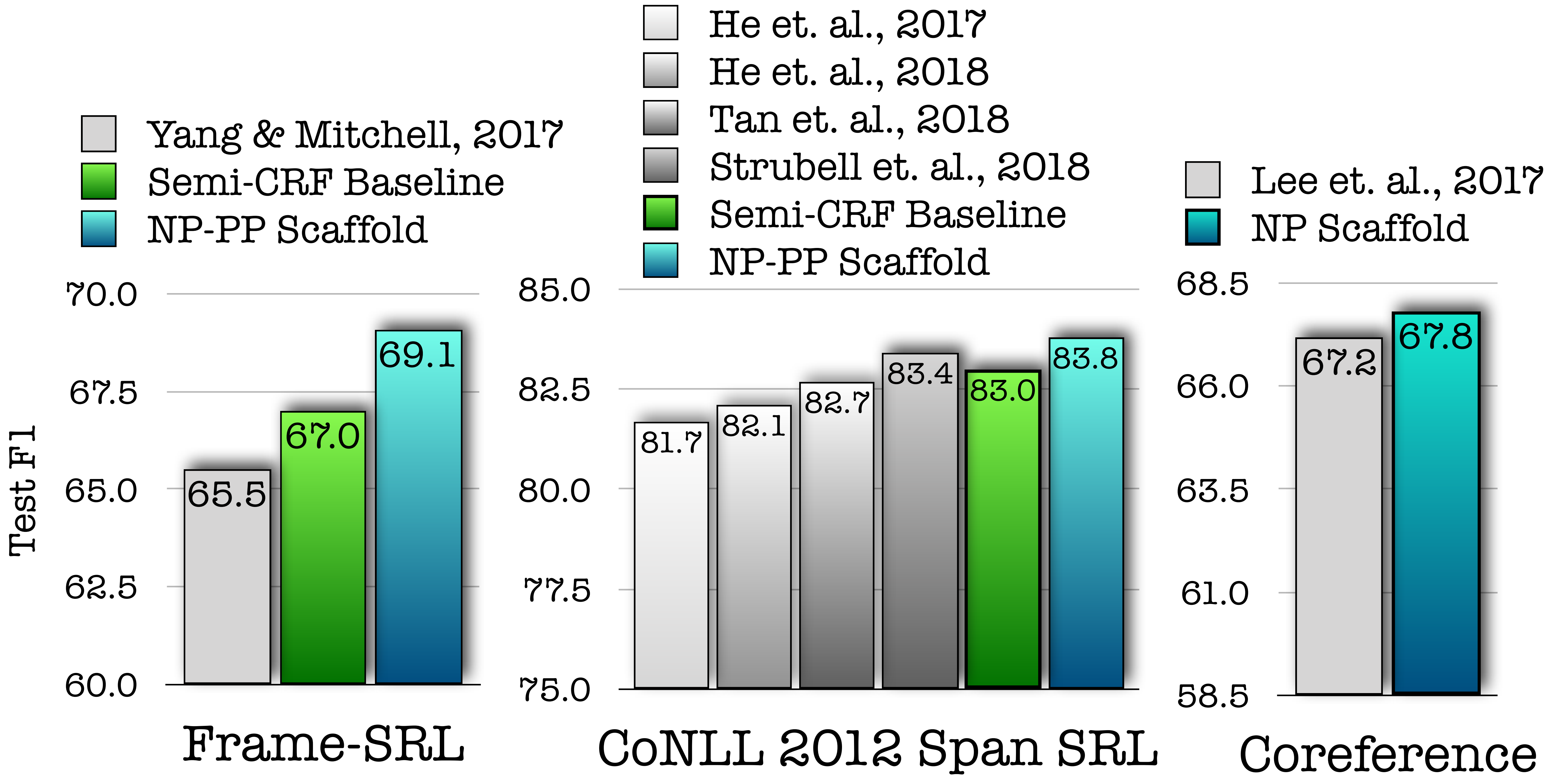
- Generalization of CRFs:
  - length of an input segment
  - in addition to its label.

$$s = \langle i, j, y_{i:j} \rangle$$

$$\Phi(\mathbf{x}, \mathbf{s}) = \sum_{k=1}^{m} \phi(s_k, x_{i_k:j_k})$$

- Training and inference given by O(ndl) dynamic programs, with a 0-order Markovian assumption.

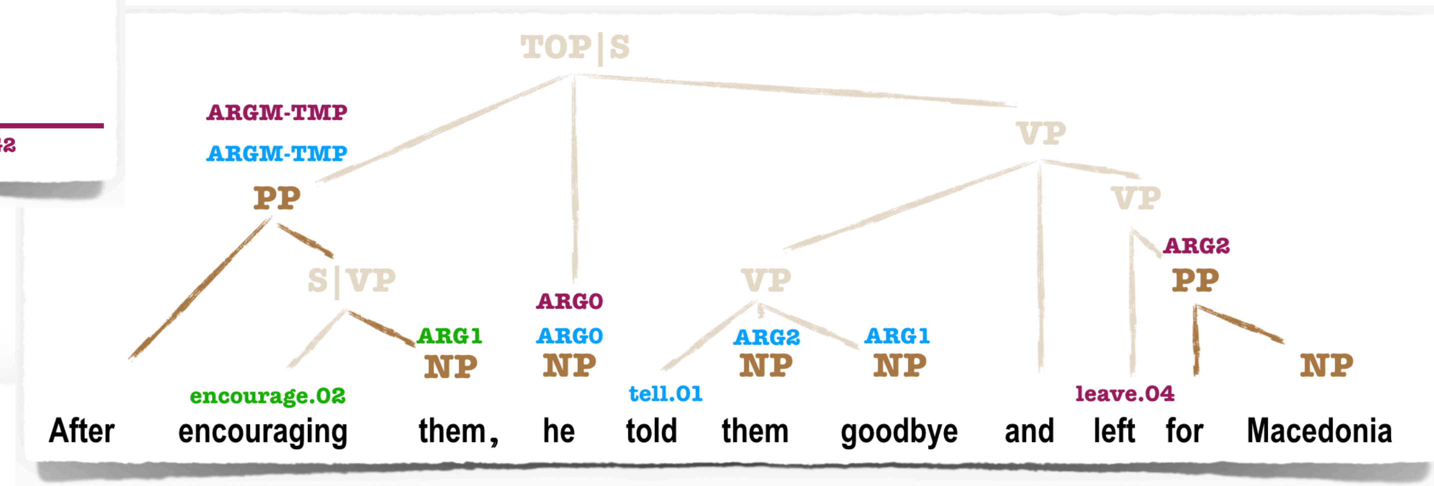
Results

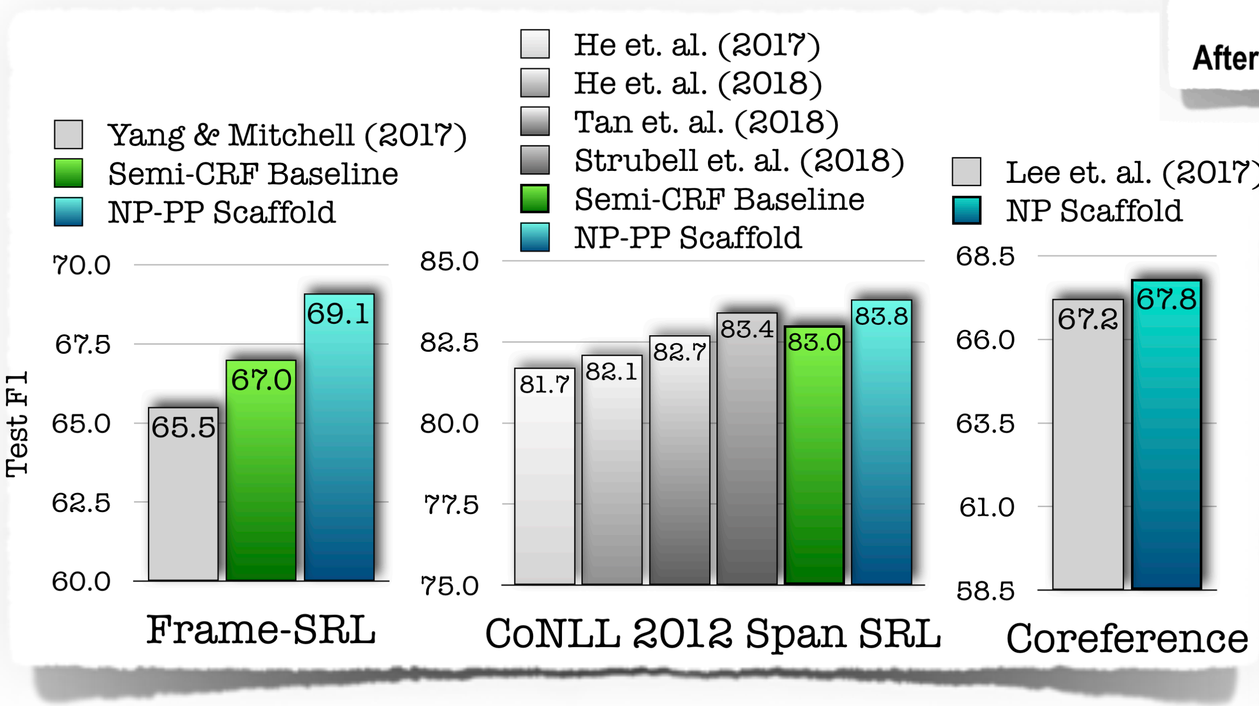


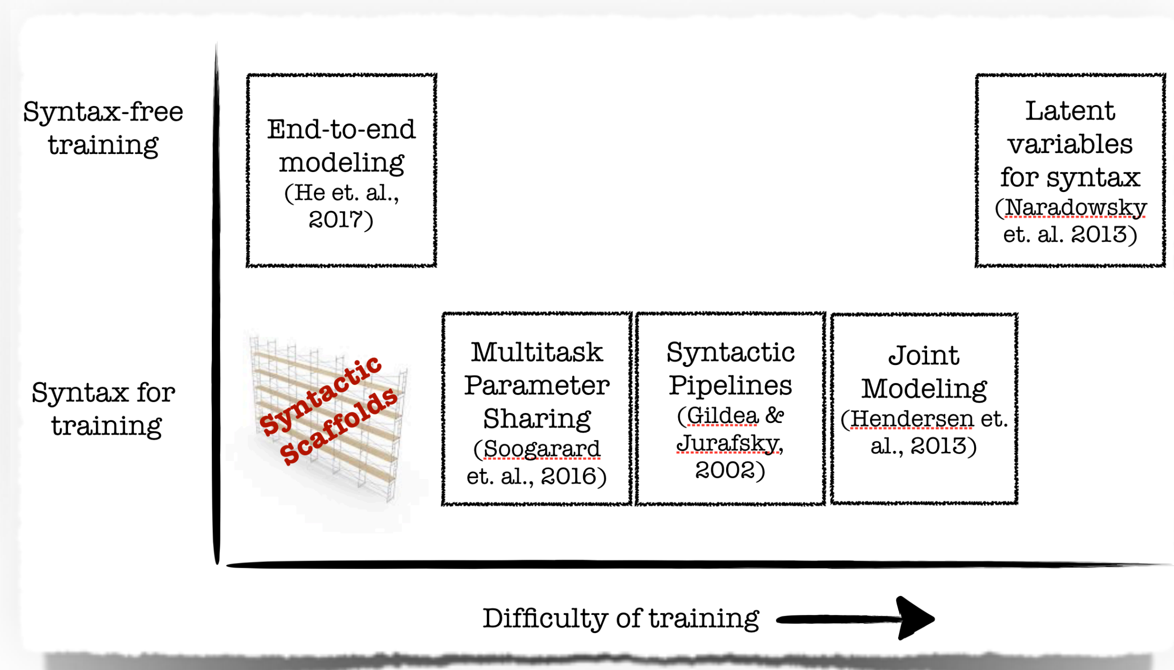
Yang & Mitchell, 2017
Semi-CRF Baseline
NP-PP Scaffold
Test F1
70.0
67.5
65.0
62.5
60.0
65.5
67.0
69.1
Frame-SRL
He et. al., 2017
He et. al., 2018
Tan et. al., 2018
Strubell et. al., 2018
Semi-CRF Baseline
NP-PP Scaffold
85.0
82.5
80.0
77.5
75.0
81.7
82.1
82.7
83.4
83.0
83.8
CoNLL 2012 Span SRL
Lee et. al., 2017
NP Scaffold
68.5
66.0
63.5
61.0
58.5
67.2
67.8
Coreference

Recap: Challenge #1

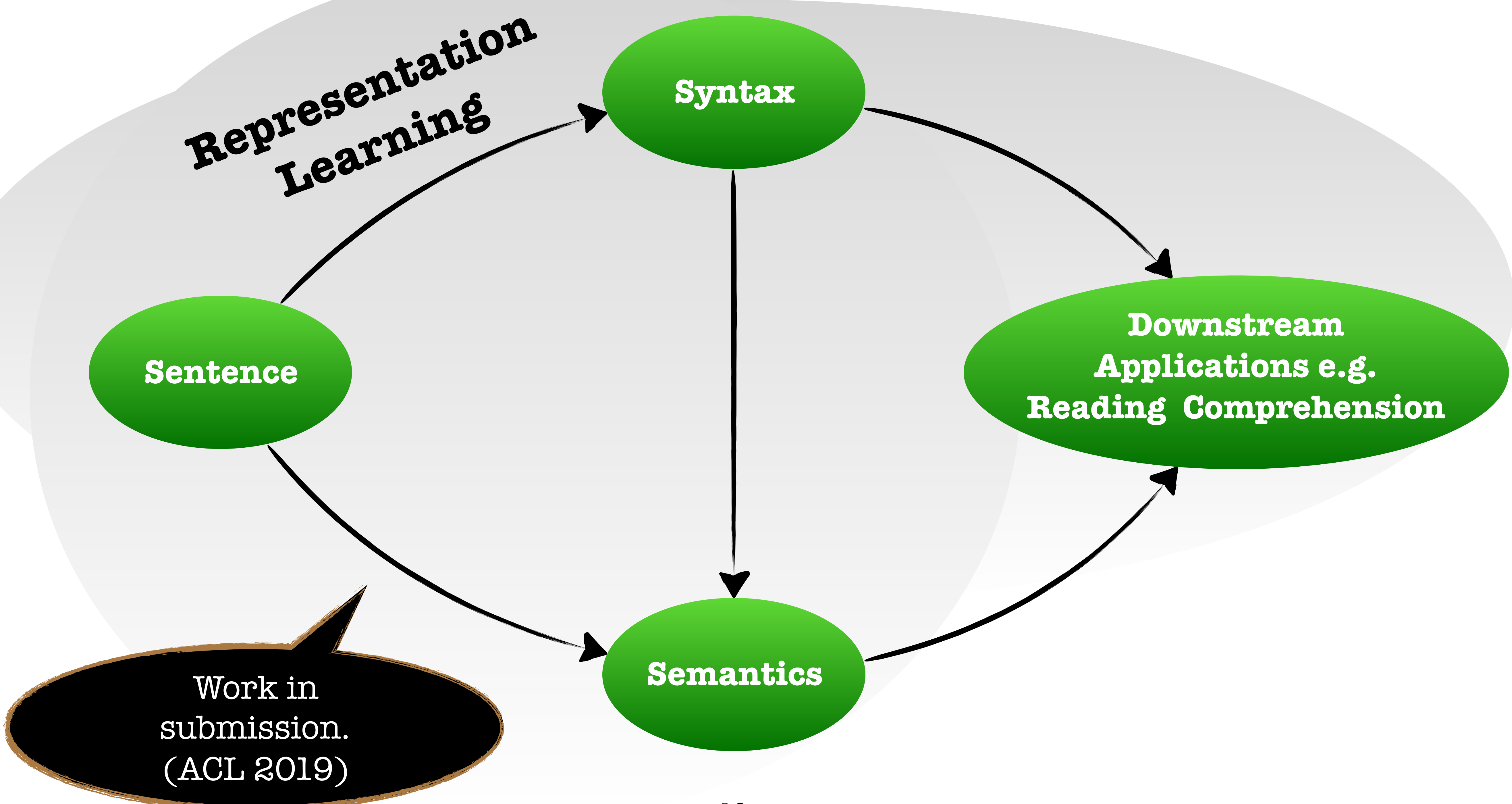
Can linguistic structure act as an informative prior for deep learning?



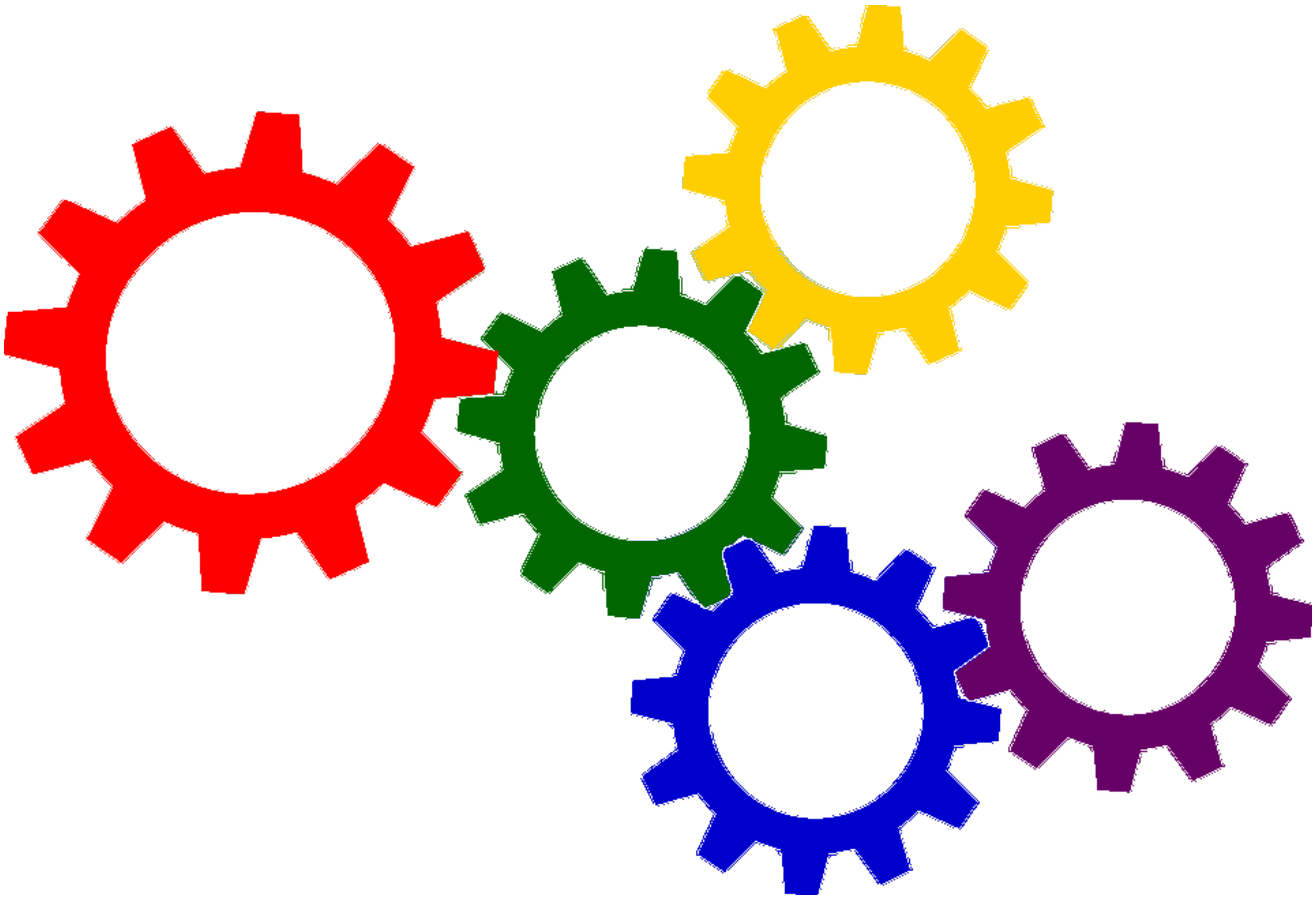




Looking ahead: Predicted Structure



Part II



Recap: BERT’s confusion

On 31 December 1687 the first organized group of Huguenots set sail from the Netherlands to the Dutch East India Company post at the Cape of Good Hope. The largest portion of the Huguenots to settle in the Cape arrived between 1688 and 1689 in seven ships as part of the organised migration, but quite a few arrived as late as **1700**; thereafter the numbers declined and only small groups arrived at a time.

The number of old Acadian colonists declined after the year **1675**.

The number of new Huguenot colonists declined after what year?

[Jia & Liang, 2017]

Percy Liang [AI Frontiers 18]

BERT [Devlin et. al., 2018]

Learning Challenges

Part I

Can linguistic structure act as an informative prior for deep learning?

- [x] Syntactic Scaffolds for Semantic Structures (EMNLP 2018)

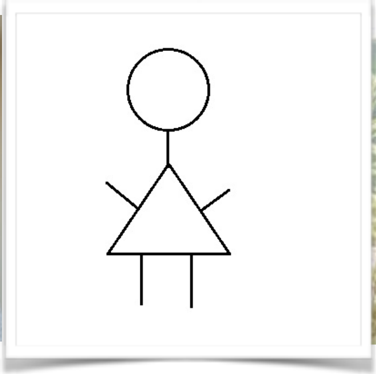
Part II

What in our data is causing models to achieve high performance?

- [ ] Annotation Artifacts in Natural Language Inference Data (NAACL 2018)

Annotation Artifacts in Natural Language Inference Data

NAACL 2018



Suchin Gururangan*
S.*
Omer Levy
Roy Schwartz
Sam Bowman
Noah A. Smith

*equal contribution

Natural Language Inference (NLI)

- Given a premise, is a hypothesis true, false or neither?

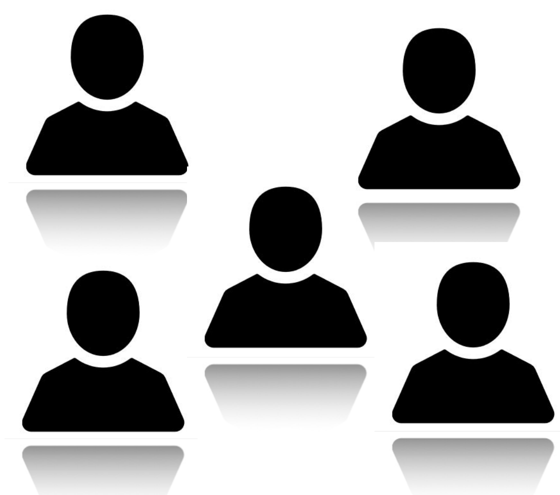
Premise	Two dogs are running through a field.
Hypothesis	The pets are sitting on a couch.

- True → **Entailment**
- False → **Contradiction**
- Cannot Say → **Neutral**

NLI Datasets

Two dogs are running through a field.

Premise



Entailment

There are animals outdoors.

Neutral

Some puppies are running to catch a stick.

Contradiction

The pets are sitting on a couch.

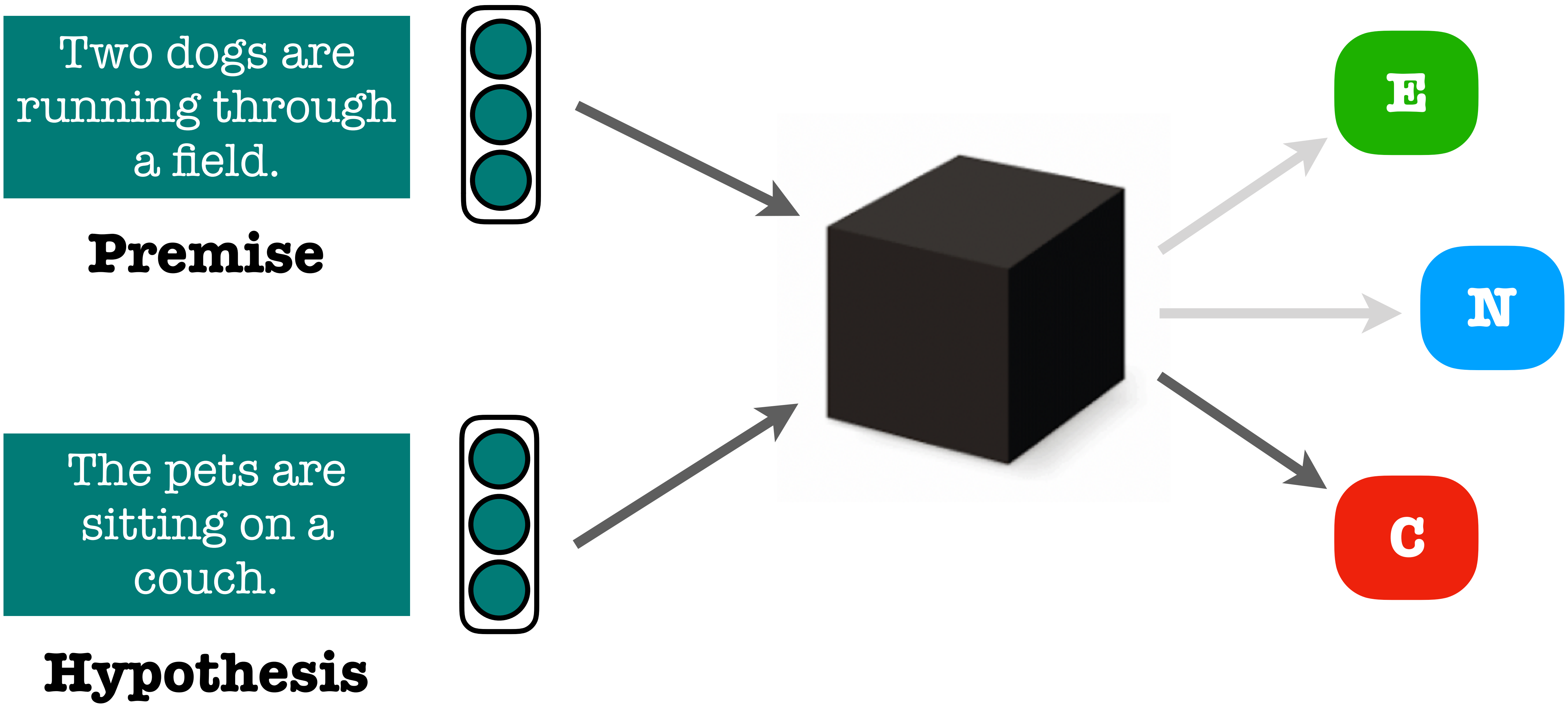
SNLI [Bowman et. al, 2015] 570 K
MultiNLI [Williams et. al., 2017] 433 K

Lots of progress

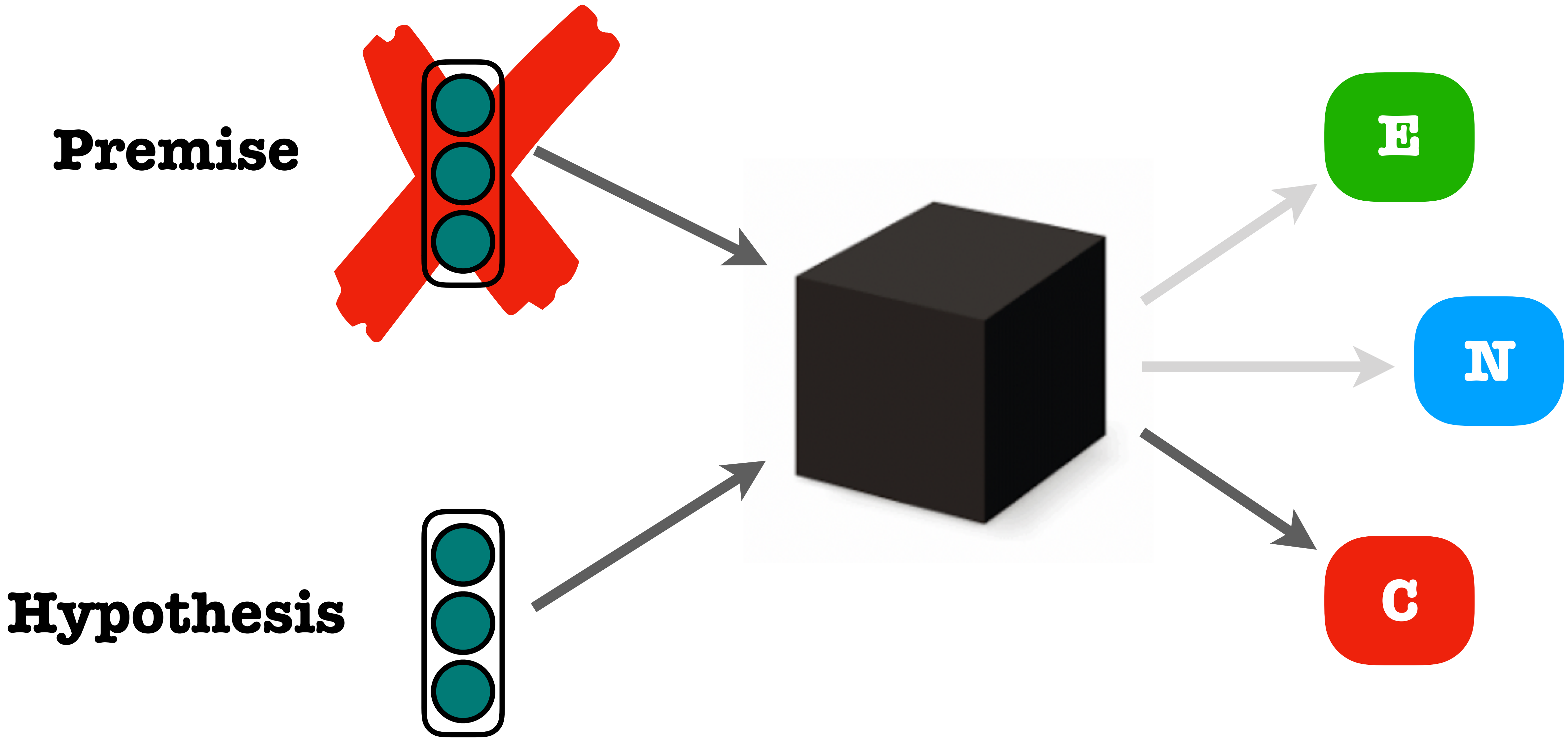
#	Team Name	Kernel	Team Members	Score	Entries	Last
1	**Allen Lao**			0.86443	4	3mo
2	**Anonymous**			0.86351	2	4mo
3	**sherry77**			0.85034	2	12d
4	**Ariel**			0.84953	10	13d
5	**ysffirst**			0.84718	6	13d
6	**ArielY**			0.84687	4	12d
7	**mattpeters**			0.84595	7	3mo
	Bidirectional LSTM			0.67507		
104	**gabrielalmeida**			0.67313	5	8mo
105	**Zippy**			0.67160	2	1y
106	**kudkudak**			0.66435	2	1y
107	**Shawn Tan**			0.65271	1	6d
	CBOW			0.65200		

MNLI Leaderboard

NLI as Text Classification

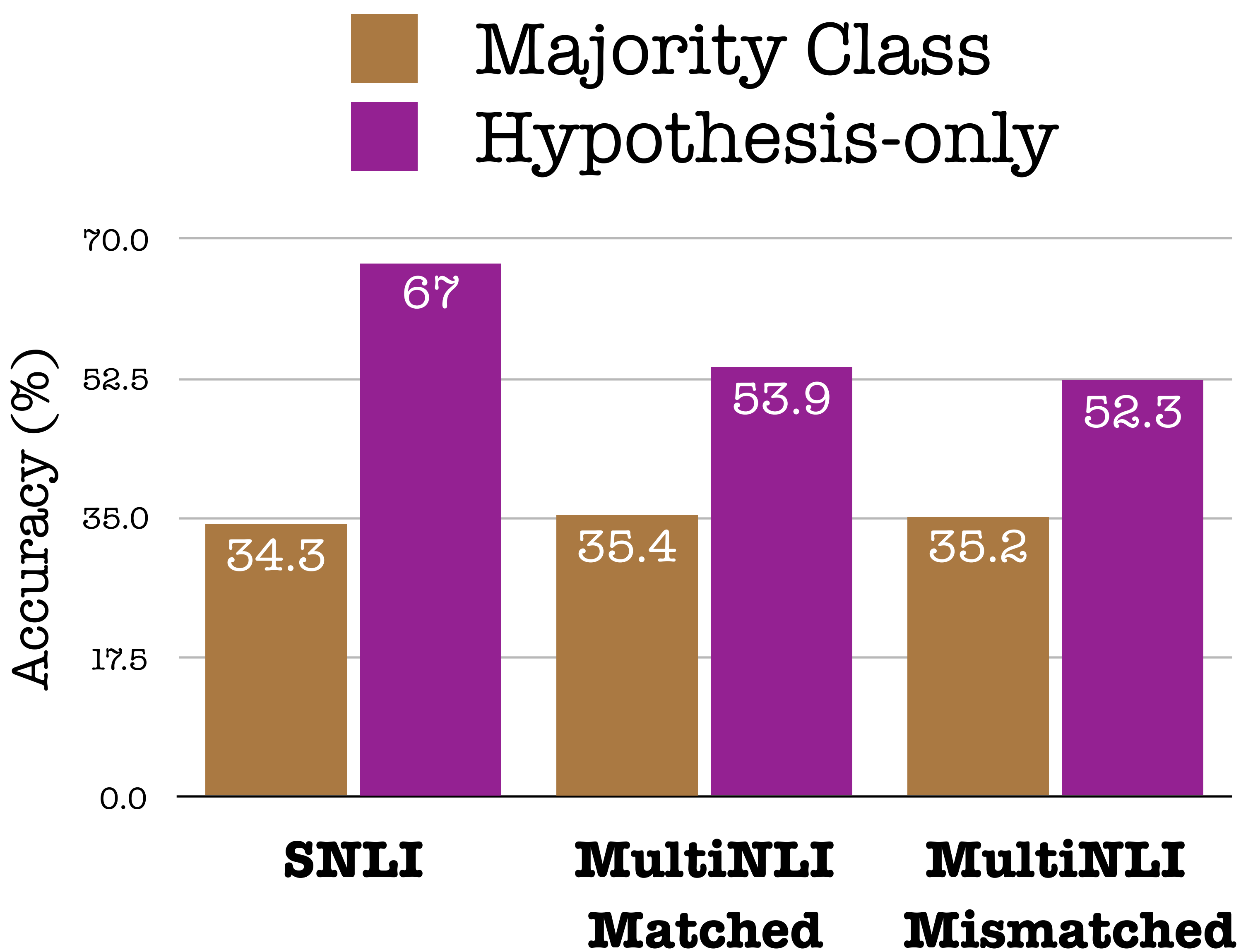


A simple experiment



fastText [Joulin et. al. 2017]

Surprising Results!



Over 50% of NLI examples can be correctly classified **without** ever observing the premise!

Entailment Artifacts

Some men and boys are playing frisbee in a grassy area.

Premise

Generalization

People play frisbee **outdoors**.

Entailment

A person in a red ~~**shirt**~~ is mowing the grass with a ~~**green**~~ riding mower.

Premise

Shortening

A person in red is cutting the grass on a riding mower.

Entailment

Neutral Artifacts

A middle-aged man works under the engine of a train on rail tracks.

Premise

Modifiers

A man is doing work on a **black** Amtrak train.

Neutral

A group of female athletes are huddled together and excited.

Premise

Purpose Clauses

They are huddled together **because** they are working together.

Neutral

Contradiction Artifacts

Older man with white hair and a red cap painting the golden gate bridge on the shore with the golden gate bridge in the distance.

Premise

Negation

Nobody wears a cap.

Contradiction

Three dogs racing on racetrack.

Premise

Cats!

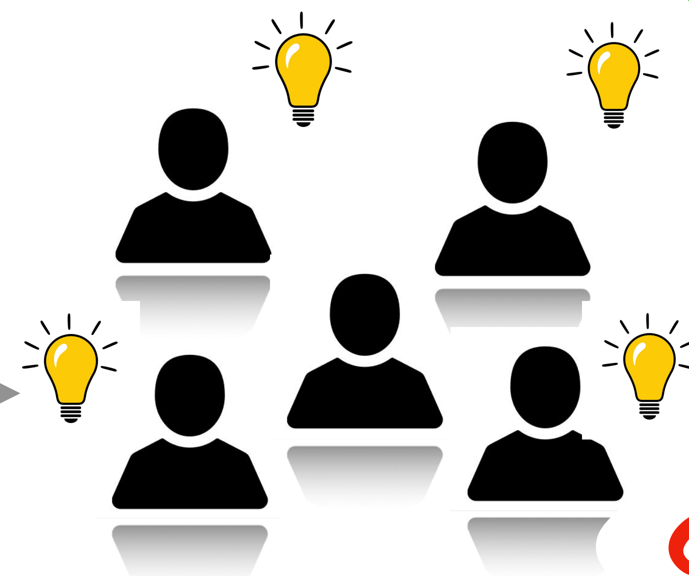
Three **cats** race on a track.

Contradiction

Annotation Artifacts

Two dogs are running through a field.

Premise



Entailment

There are **animals** outdoors.

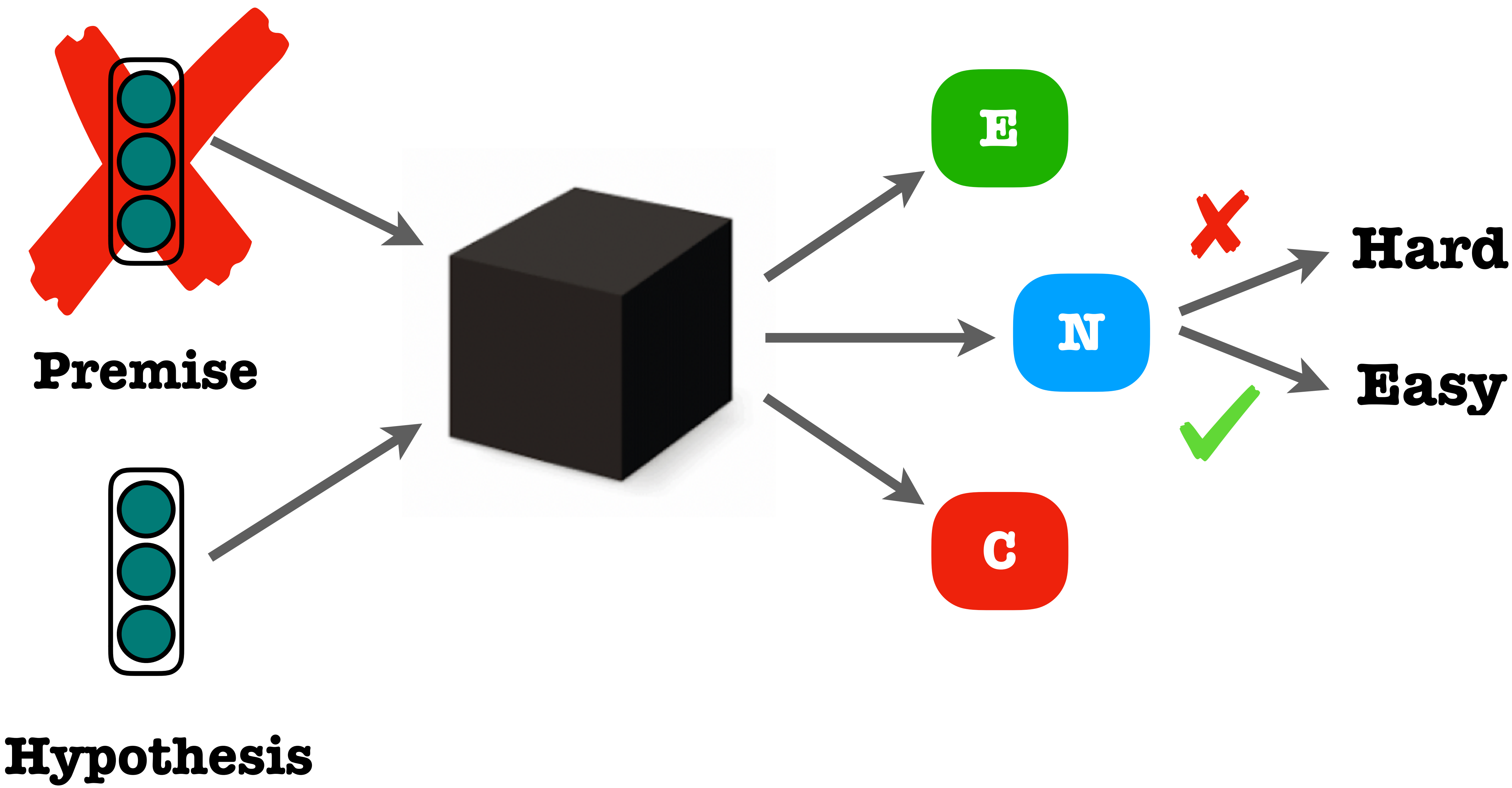
Neutral

Some puppies are running **to catch a stick**.

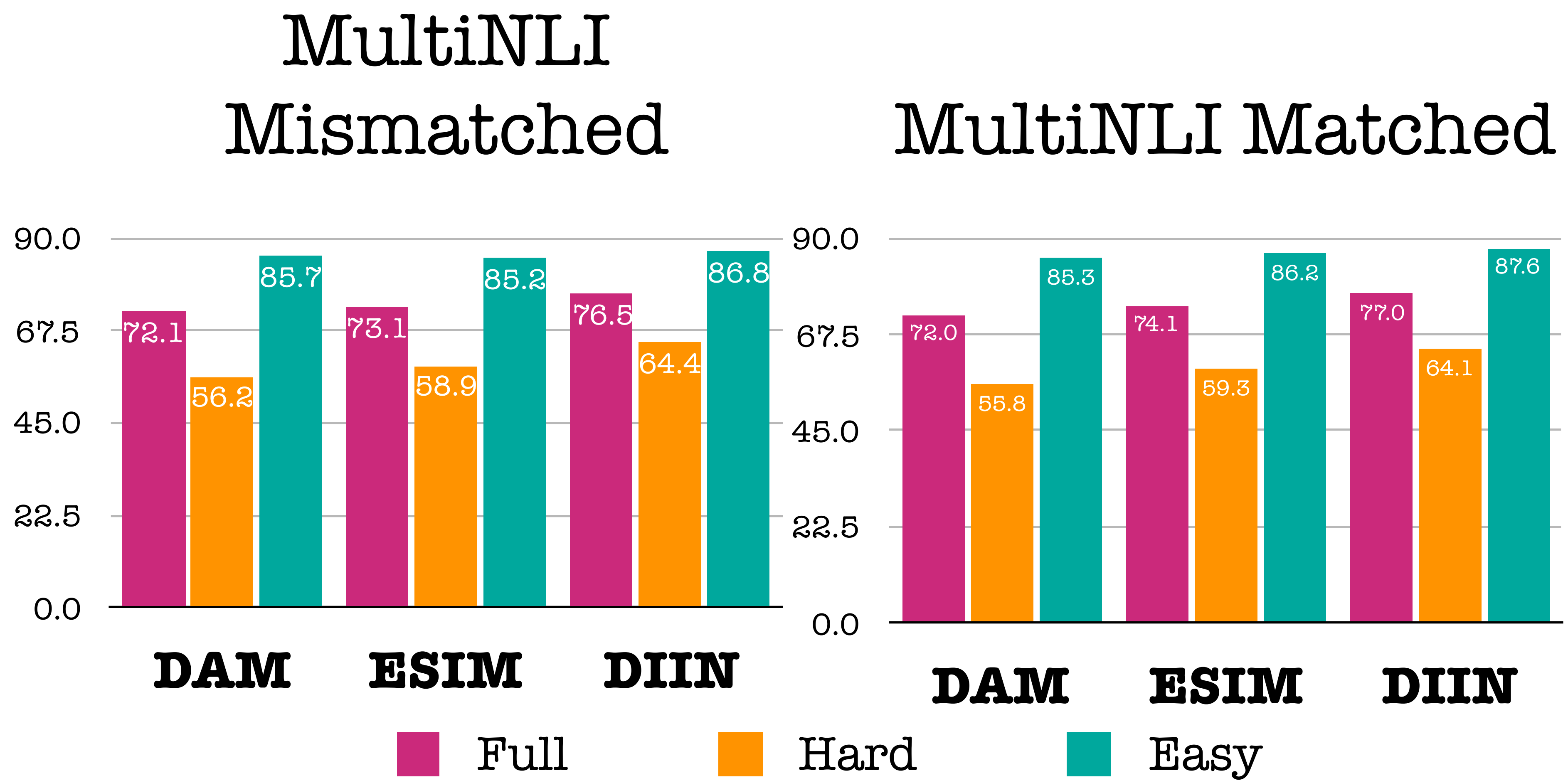
Contradiction

The pets are sitting on a couch.

Can we filter out examples with artifacts?

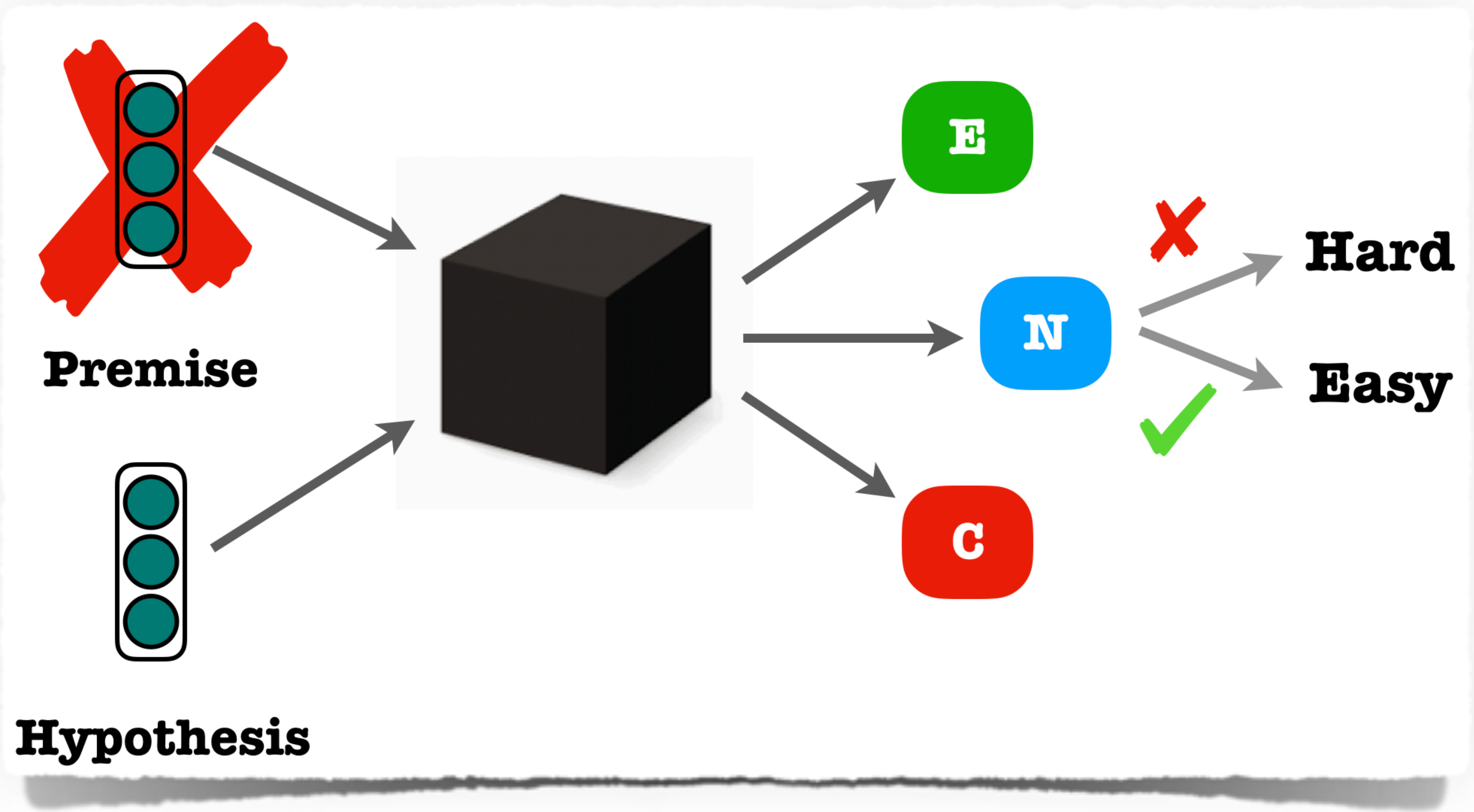


Revisiting NLI models



DAM - Decomposable Attention Model [Parikh et. al. 2016]
ESIM - Enhanced Sequential Inference Model [Chen et. al., 2017]
DIIN - Densely Interactive Inference Network [Gong et. al. 2018]

Revisit: Can we filter out examples with artifacts?



- Hard examples exhibit their own artifacts!
- Artifacts are still valid examples...

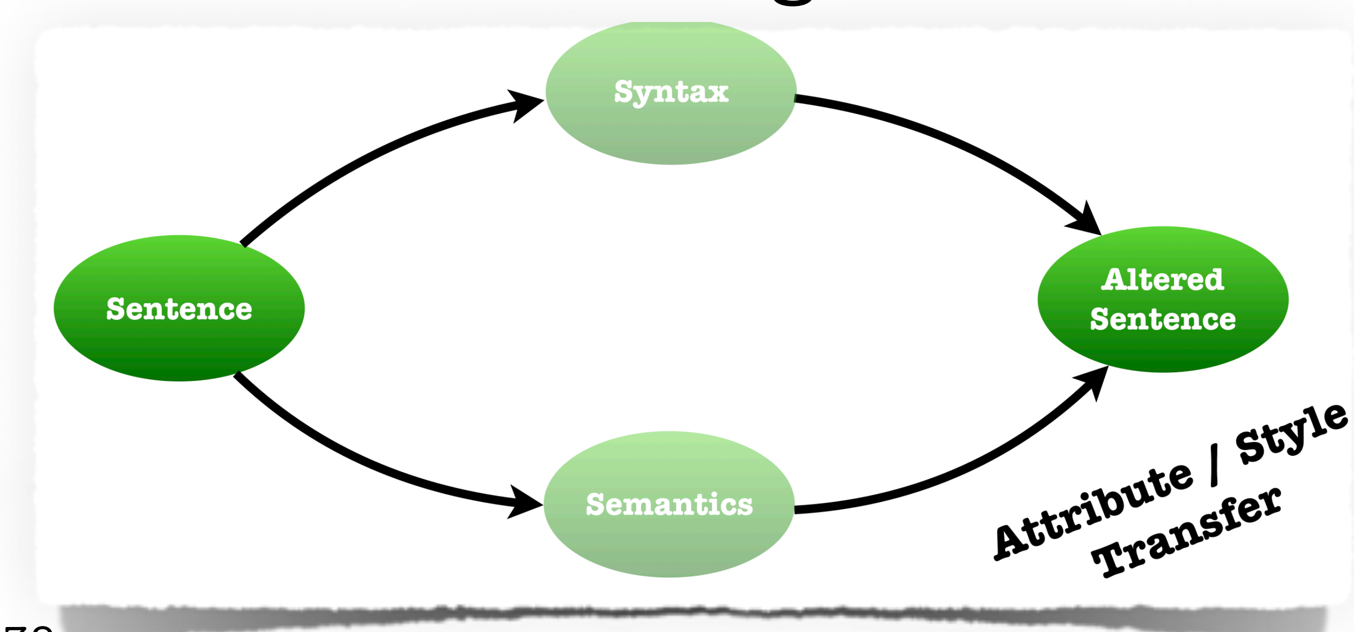
Looking ahead:
Learning from Datasets with Artifacts

- Intuition:
  - Models which exploit artifacts == models which can detect artifacts.
  - Stylistic global features.
- Subsampling large datasets → weight each example based on how hard it could be [Beygelzimer et. al., 2015, Coleman et. al., 2018].

Easy  **Hard**

Looking Ahead: Improved Data Collection

- Annotation Instruction: Avoid simple heuristics!
- Real-time rewards and batch reviews of annotated examples. E.g. SWAG [Zellers et. al., 2018], DROP [Dua et. al., 2019]
- Alternatives to human elicitation for building datasets?

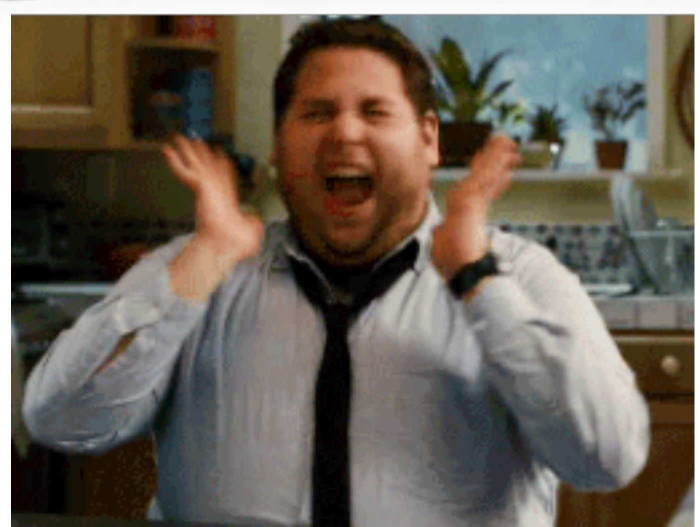


In conclusion : It's an exciting time!

The New York Times

Finally, a Machine That Can Finish Your Sentence

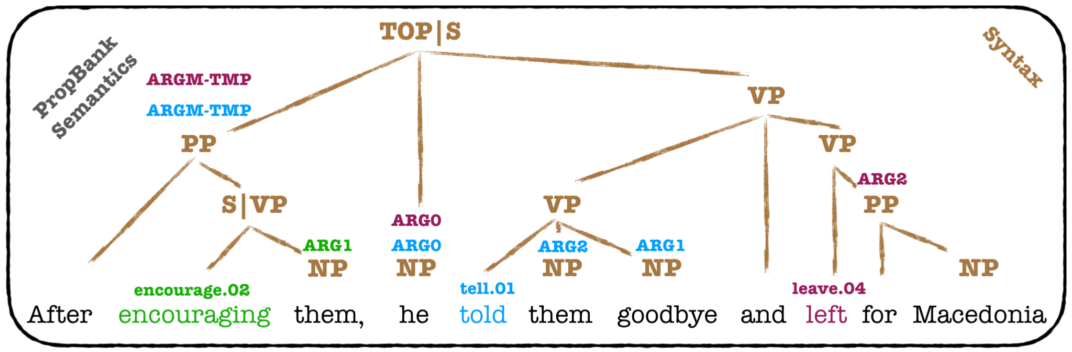
Completing someone else's thought is not an easy trick for A.I. But new systems are starting to crack the code of natural language.



In conclusion - Learning Challenges

Part I

Can linguistic structure act as an informative prior for deep learning?



Predicted structure can help representation learning.

Part II

What in our data is causing models to achieve high performance?

Need models robust to artifacts.

Thanks!

http://www.cs.cmu.edu/~sswayamd

swabhs

swabhz