Character Eyes: Seeing Language through Character-Level Taggers

Yuval Pinter
@yuvalpi

Marc Marone
@ruyimarone

Jacob Eisenstein
@jacobeisenstein

Bloomberg

Microsoft Research

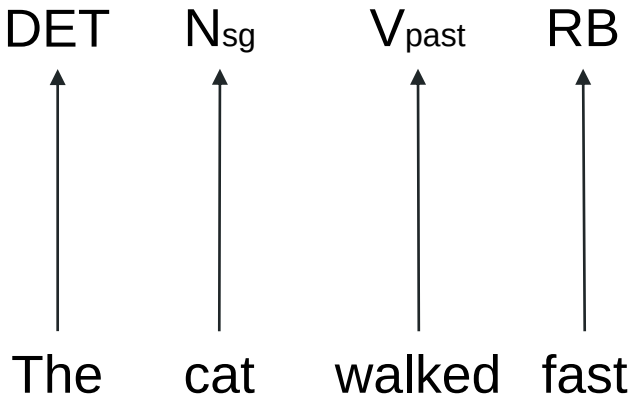
Google

Georgia Institute of Technology

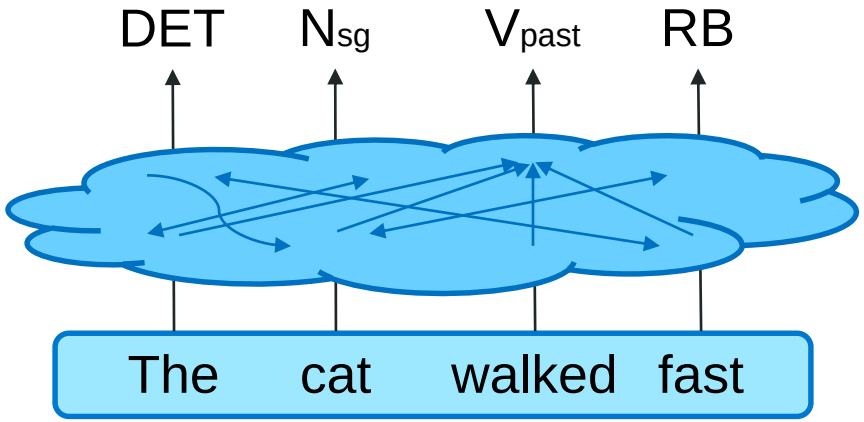
Blackbox NLP 2019

https://github.com/ruyimarone/character-eyes

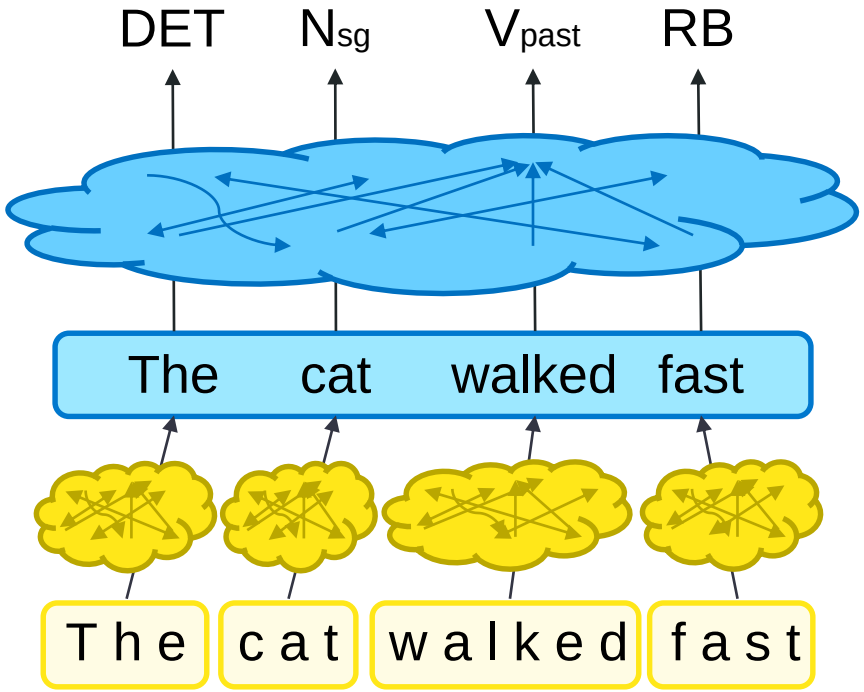
Taggers



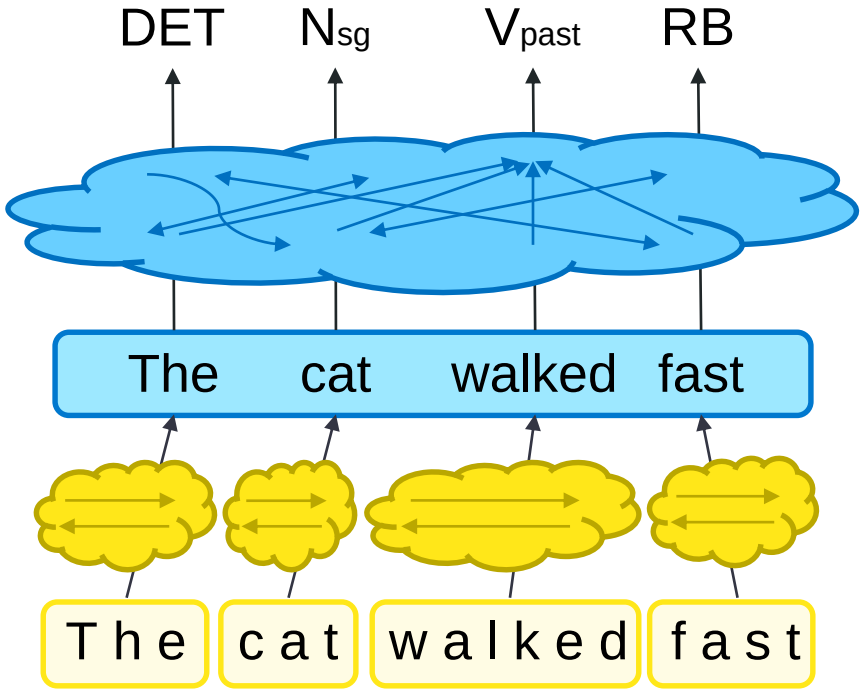
Neural Taggers



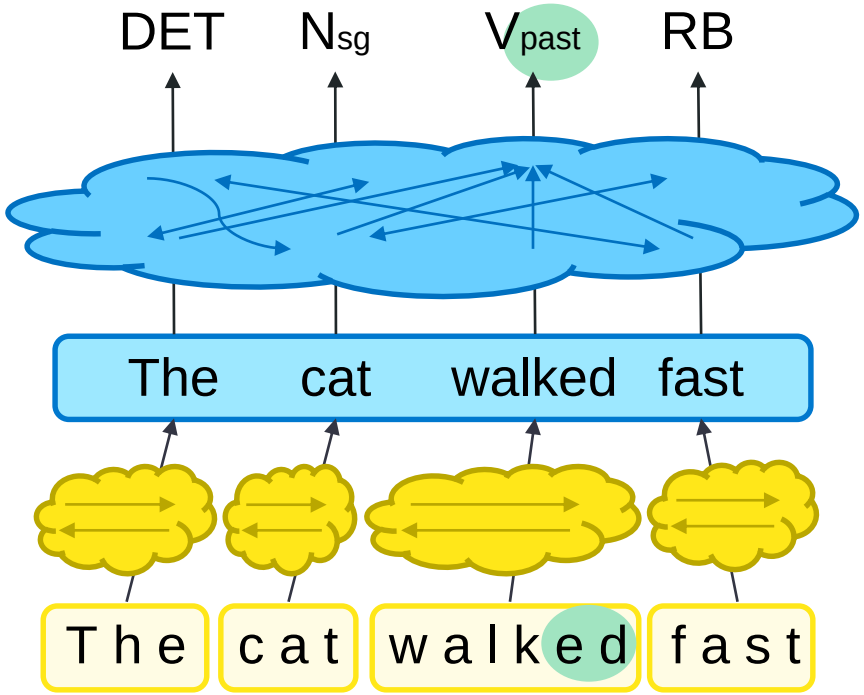
Character-level Neural Taggers



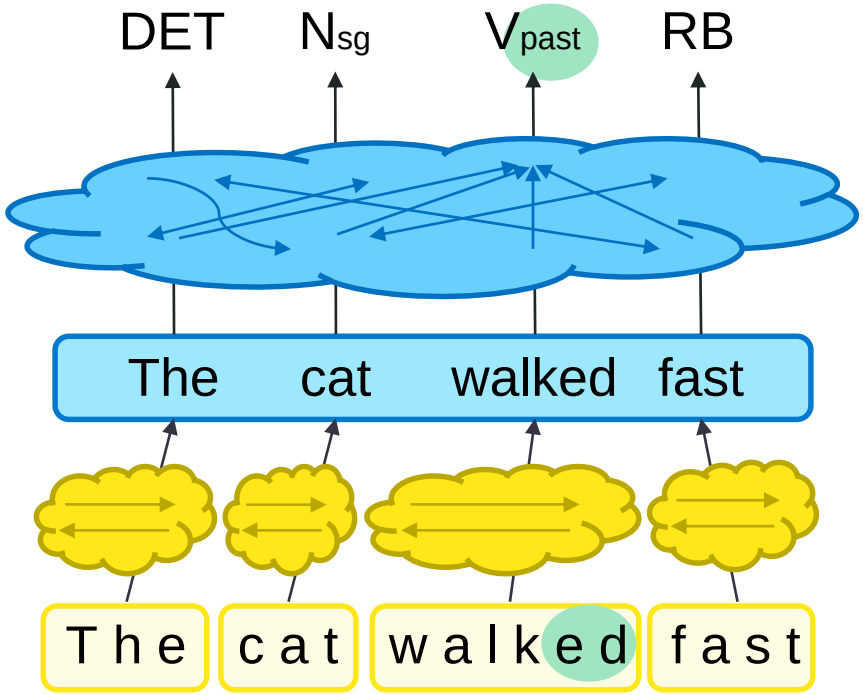
Character-level Recurrent Neural Taggers



Recurrent Taggers – Good at Finding Morphemes?

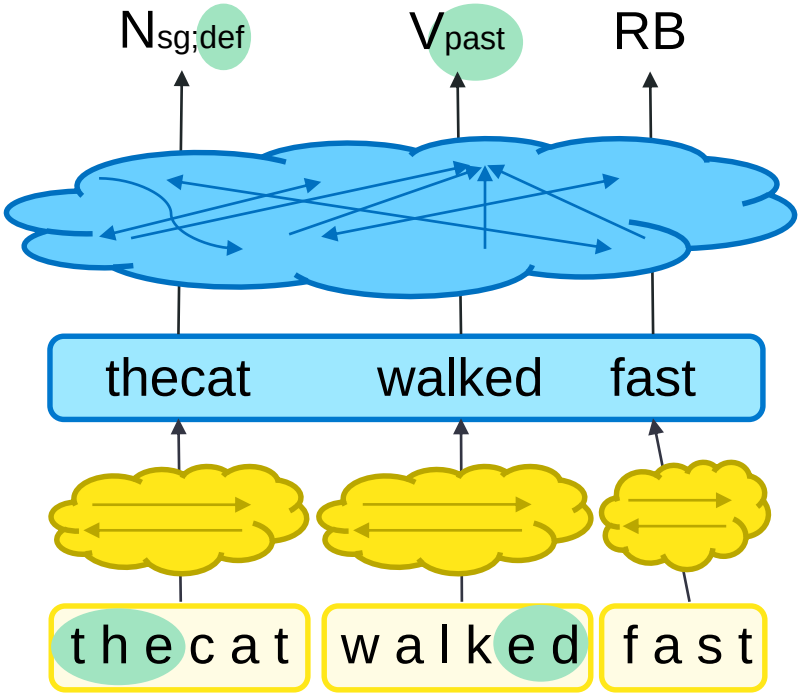


Recurrent Taggers – Good at Finding Morphemes?

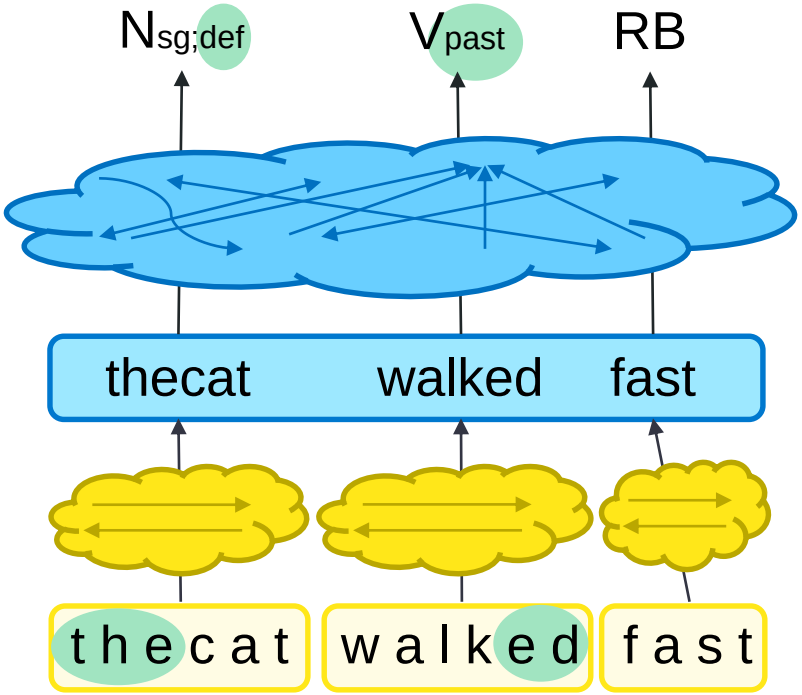


Agglutination

Recurrent Taggers – Good at Prefixes **and** Suffixes?

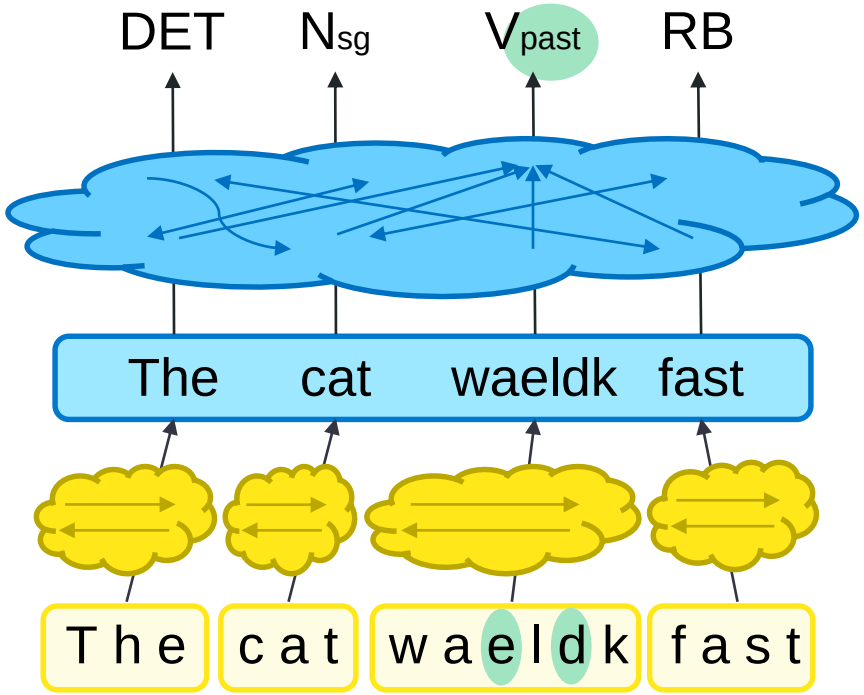


Recurrent Taggers – Good at Prefixes **and** Suffixes?

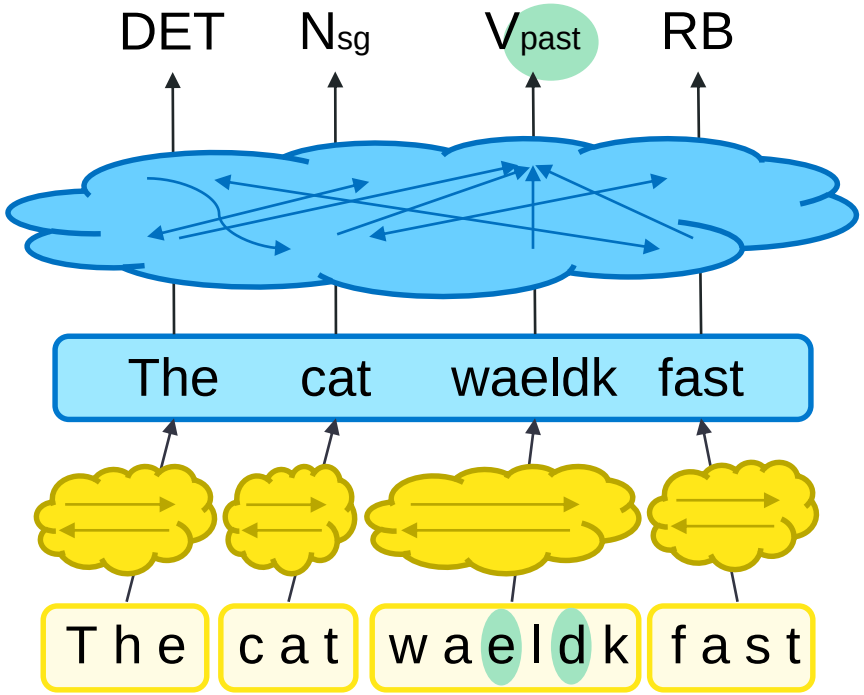


Prefixing morphology
(e.g. Coptic)

Recurrent Taggers – Can They Handle diSCoNtinUiTY?

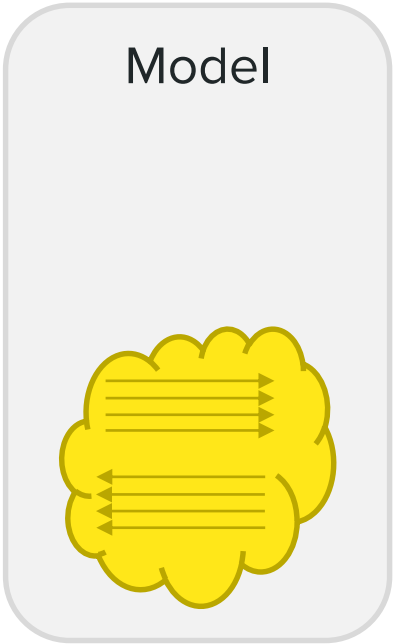


Recurrent Taggers – Can They Handle diSCoNtinUiTY?

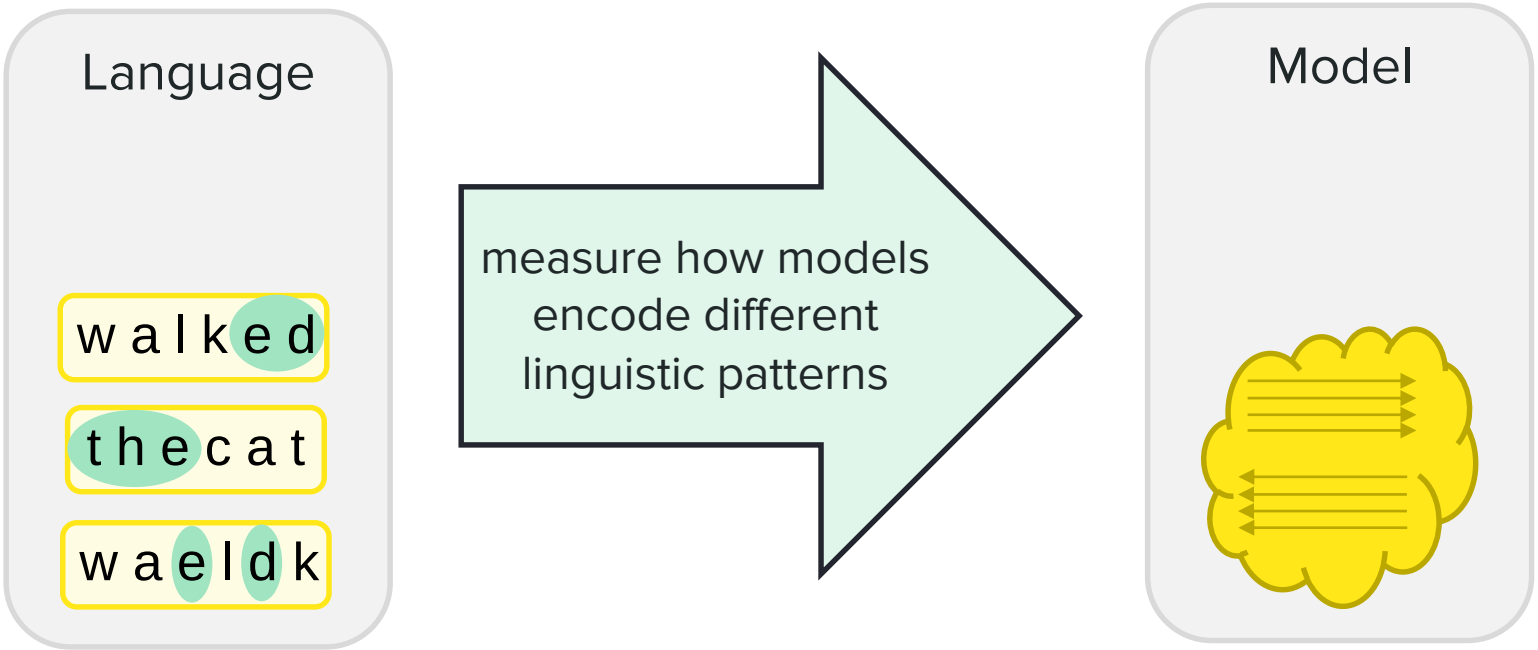


Introflexive morphology
(Hebrew, Arabic)

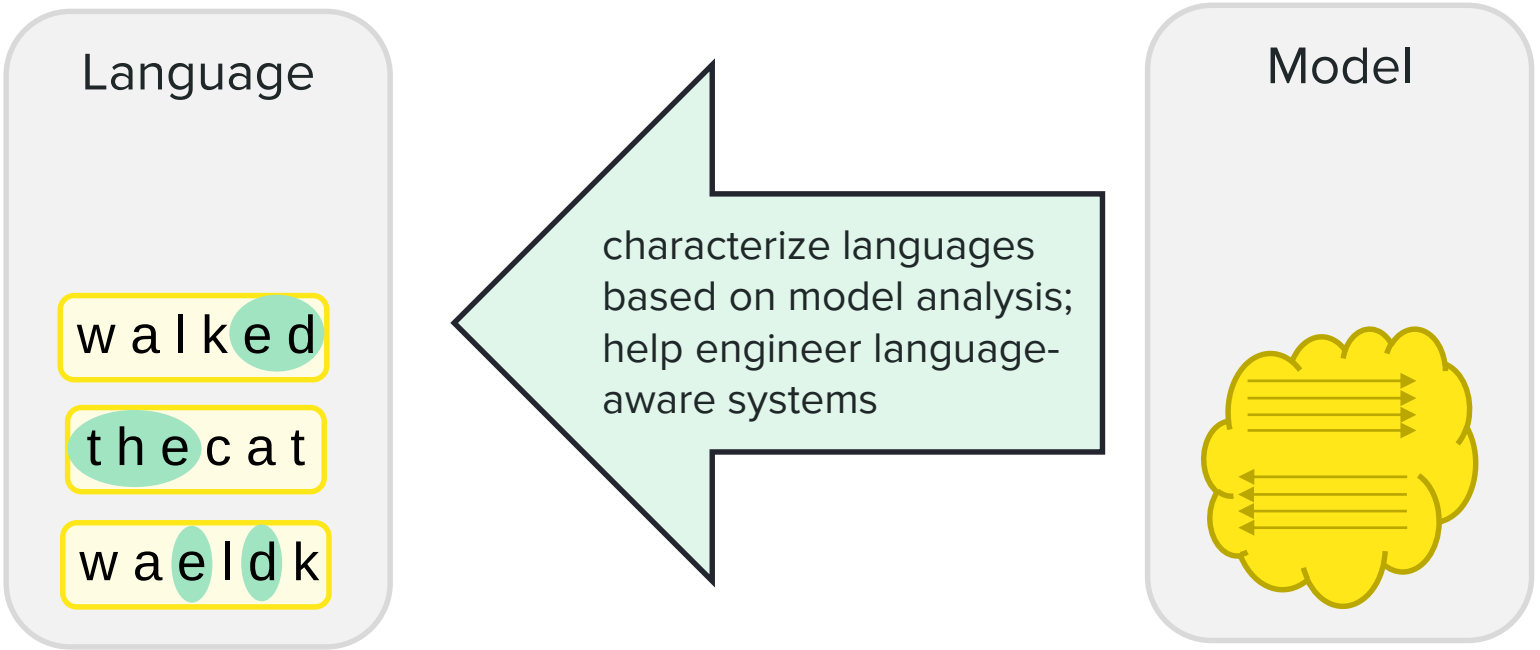
Main Idea(s)



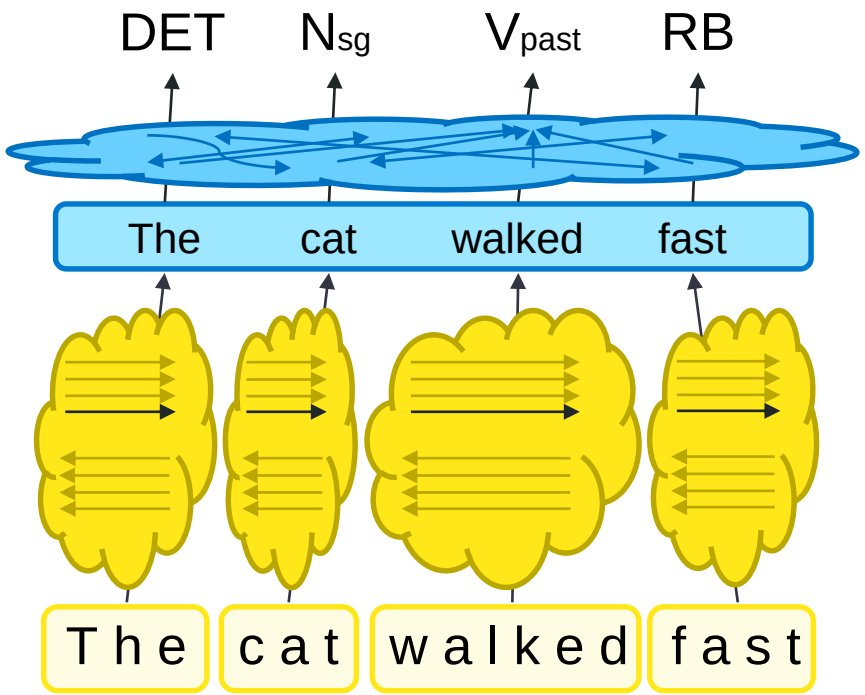
Main Idea(s)


Language
w a l k e d
t h e c a t
w a e l d k
measure how models encode different linguistic patterns
Model

Main Idea(s)

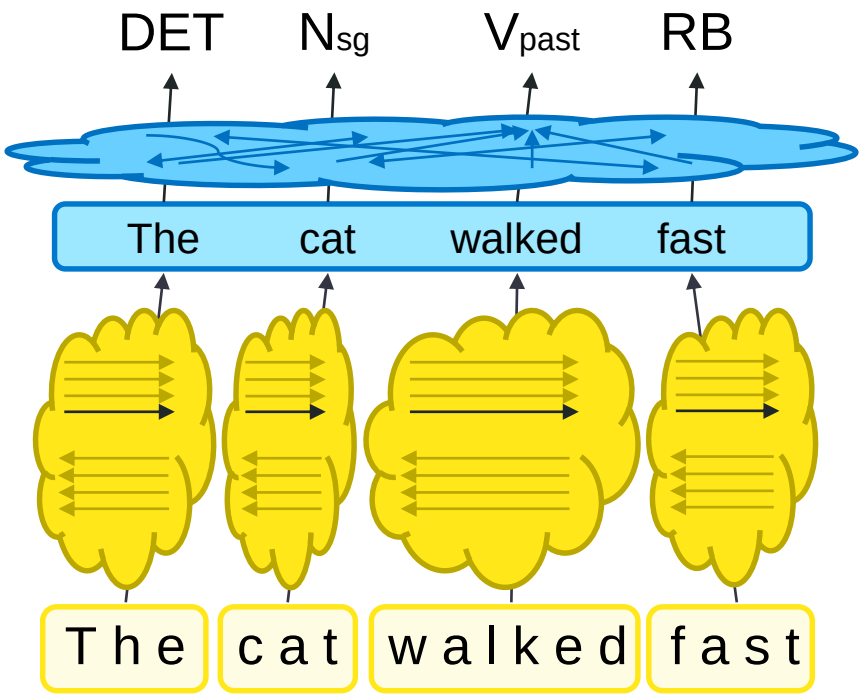

Language
w a l k e d
t h e c a t
w a e l d k
characterize languages based on model analysis; help engineer language-aware systems
Model

Analysis Primitive – Unit Decomposition



Analysis Primitive – Unit Decomposition

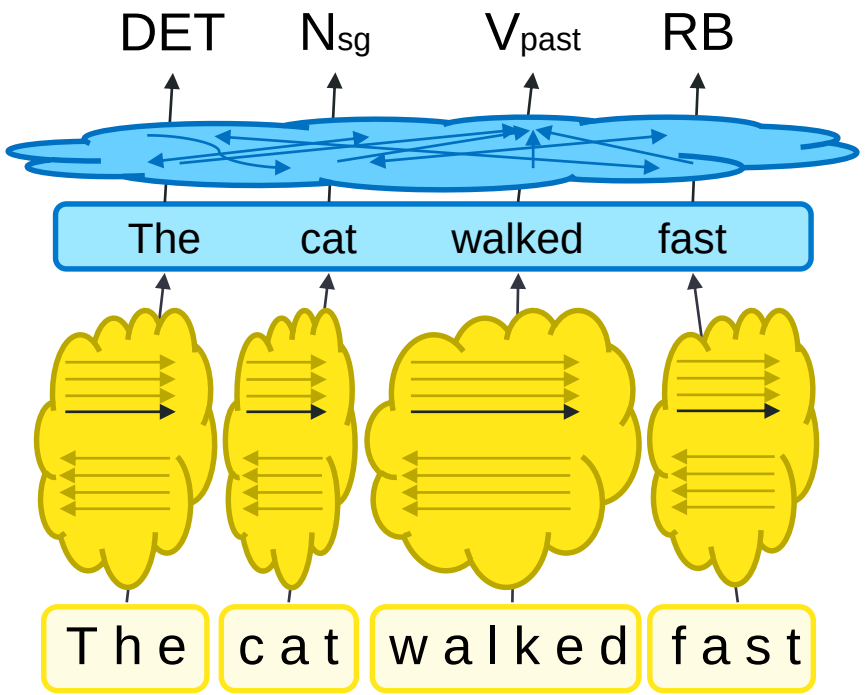
- Assumption: units are “in charge” of tracking morphemes that help predict POS



Hidden unit #n

Analysis Primitive – Unit Decomposition

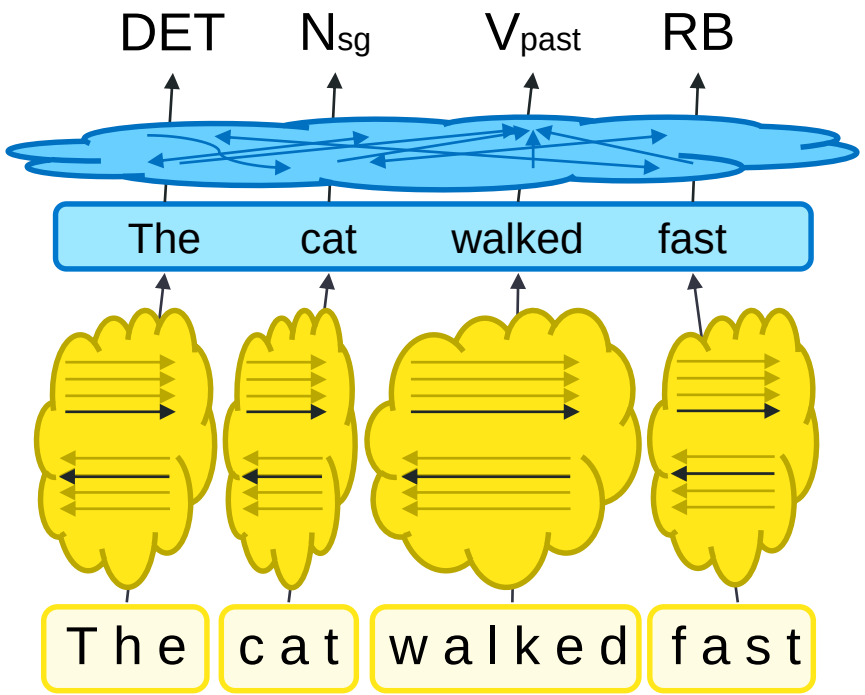
- Assumption: units are “in charge” of tracking morphemes that help predict POS
- Hypothesis: easy for **agglutinations**, difficult for **introflexions**



Hidden unit #n

Analysis Primitive – Unit Decomposition

- Assumption: units are “in charge” of tracking morphemes that help predict POS
- Hypothesis: easy for **agglutinations**, difficult for **introflexions**
- Hypothesis: unit’s **direction** affects ease of tracking **suffixes** vs. **prefixes**



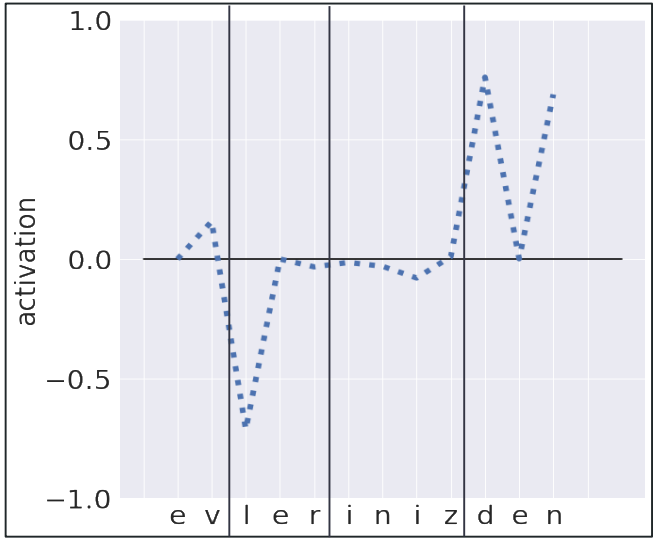
Evidence?

- Turkish is an **agglutinative** language
  - *ev* ‘house’; *ev**ler*** ‘houses’; *ev**ler**iniz*** ‘your houses’; *ev**ler**iniz**den*** ‘from your houses’

Evidence?

- Turkish is an **agglutinative** language
  - *ev* ‘house’; *ev**ler*** ‘houses’; *ev**ler**iniz* ‘your houses’; *ev**ler**iniz**den*** ‘from your houses’

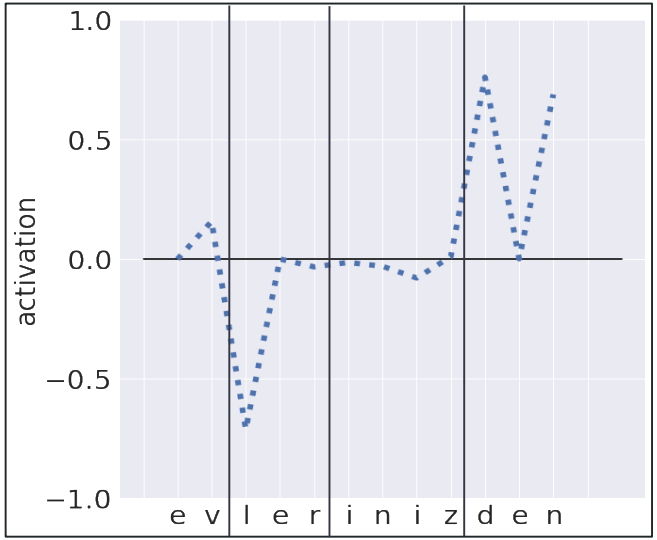
Unit 3 (→)

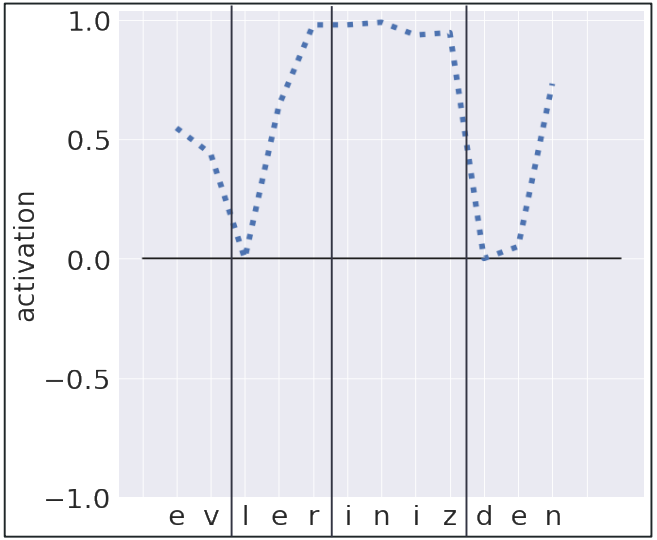


Evidence?

- Turkish is an **agglutinative** language
  - *ev* ‘house’; *ev**ler*** ‘houses’; *ev**ler**iniz* ‘your houses’; *ev**ler**iniz**den*** ‘from your houses’

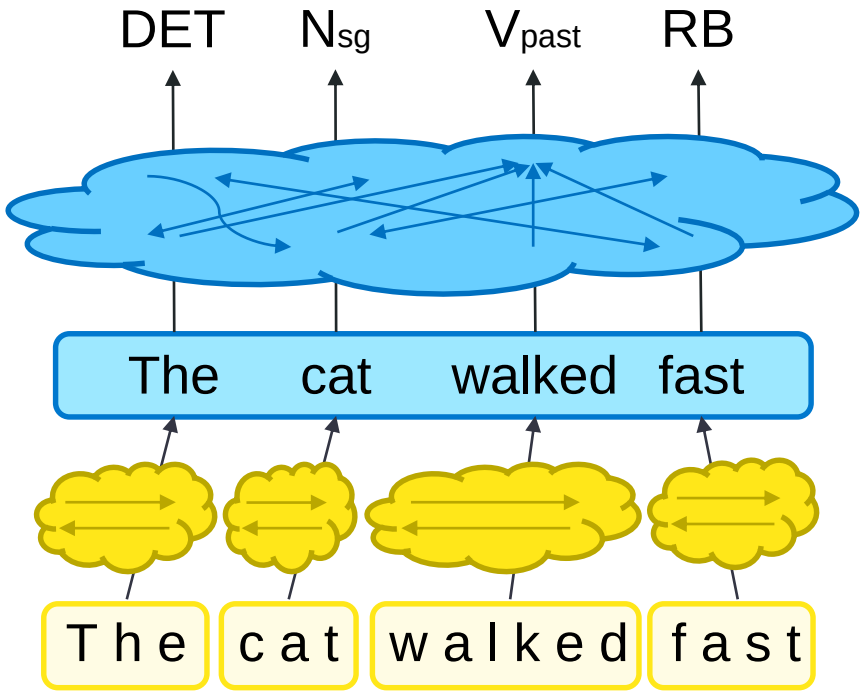
Unit 3 (→)





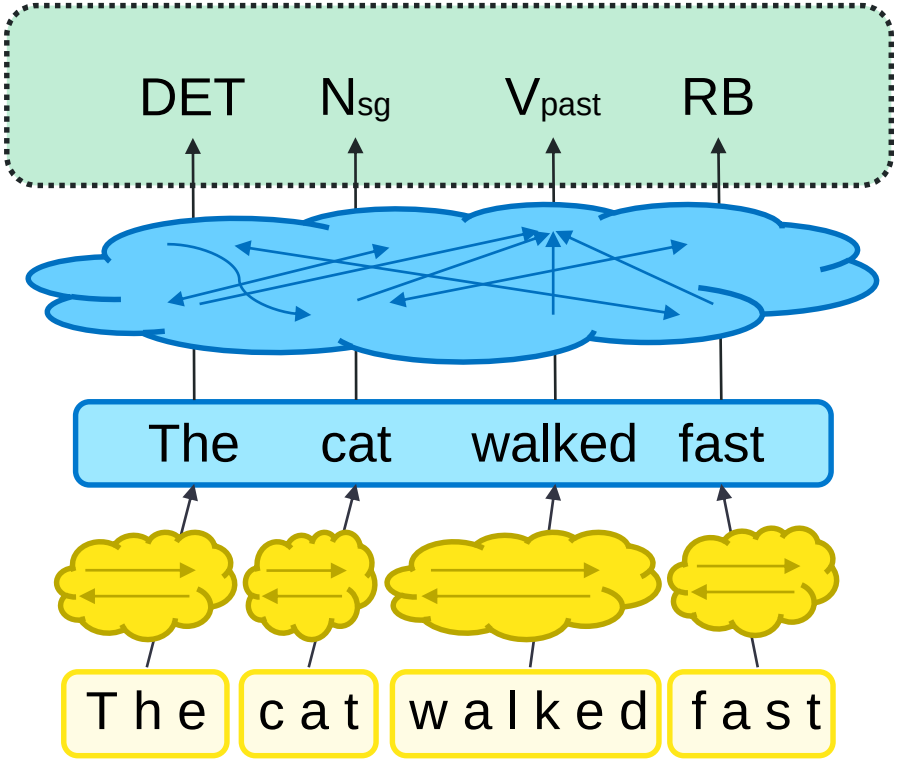
Unit 124 (←)

Model & Data



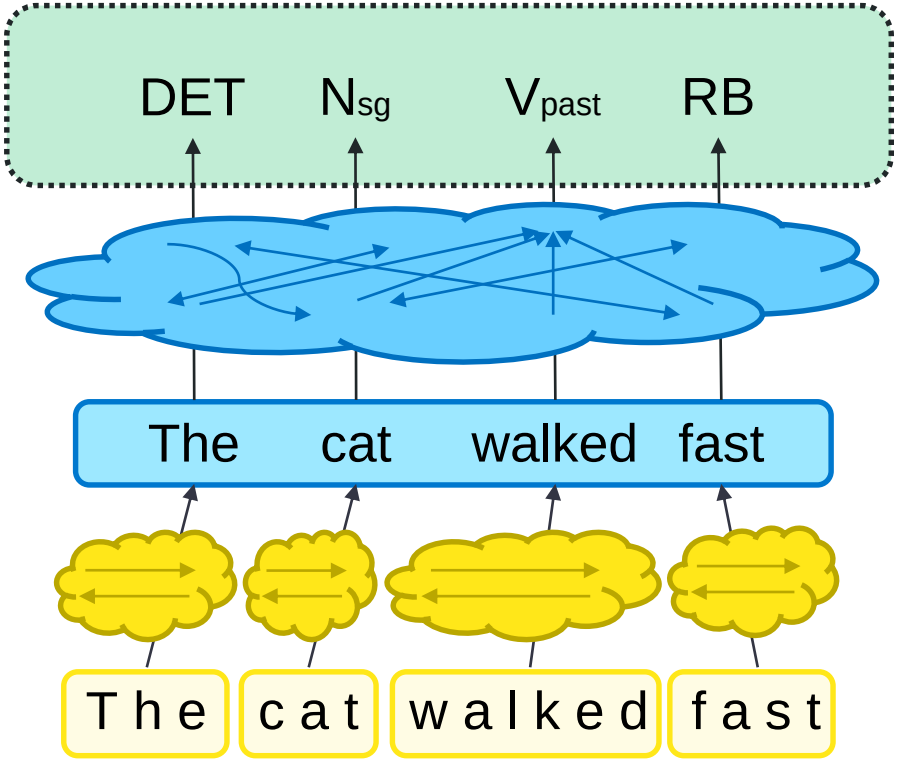
Model & Data

- Universal Dependencies (n=24)
  - POS tags + Morphosyntactic Descriptions



Model & Data

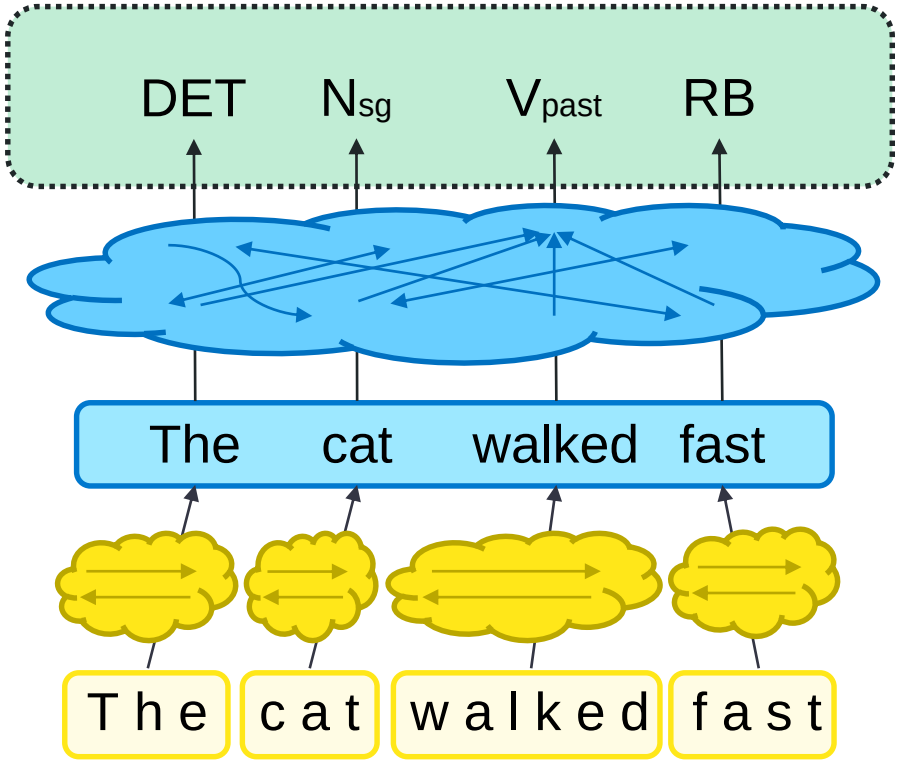
- Universal Dependencies (n=24)
  - POS tags + Morphosyntactic Descriptions
- Linguistic diversity – morph. synthesis:
  - 5 agglutinative languages
  - 2 introflexive languages
  - 3 isolating, 14 fusional



Source for language classes: WALS

Model & Data

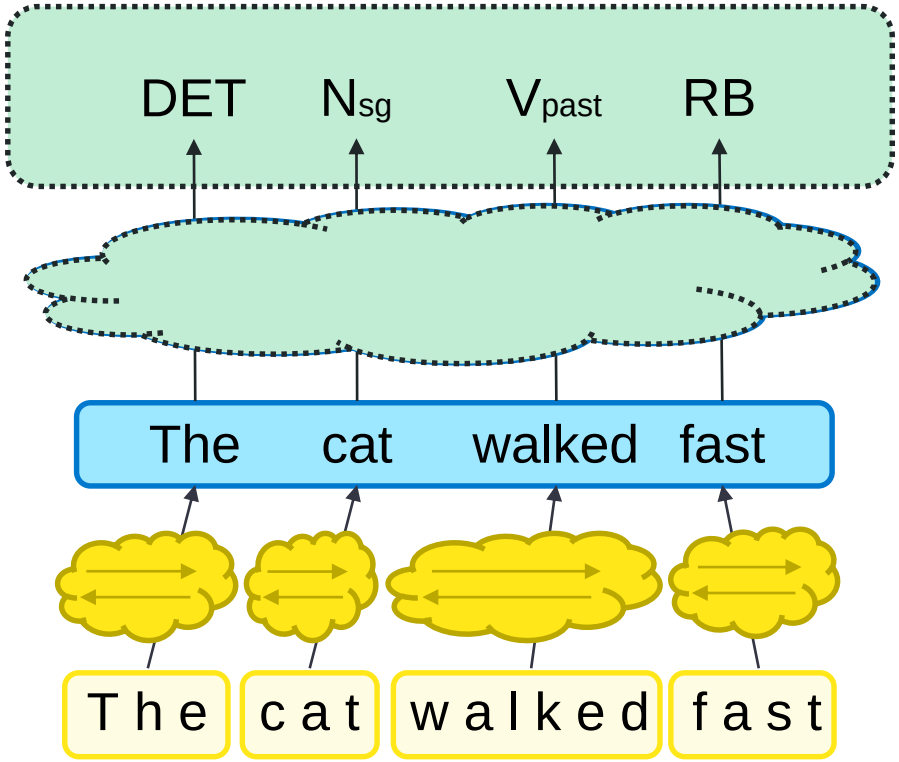
- Universal Dependencies (n=24)
  - POS tags + Morphosyntactic Descriptions
- Linguistic diversity – morph. synthesis:
  - 5 agglutinative languages
  - 2 introflexive languages
  - 3 isolating, 14 fusional
- Linguistic diversity – affixation:
  - (All) 1 prefixing language
  - 2 non-affixing
  - 2 equally pre- and suffixing
  - 19 suffixing



Source for language classes: WALS

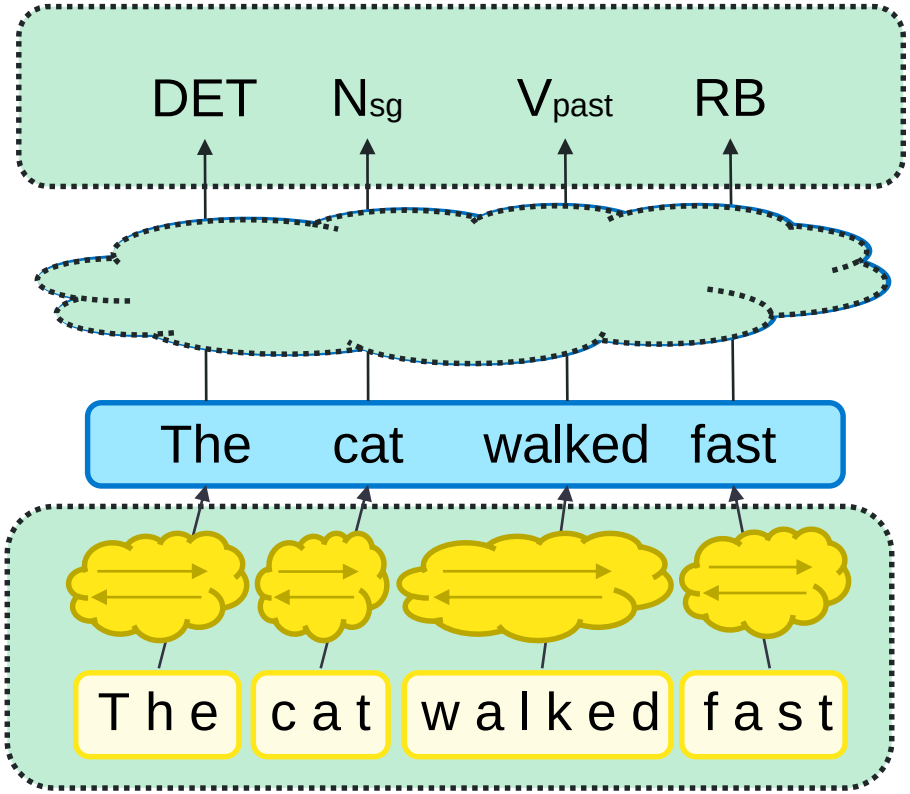
Model & Data

- Universal Dependencies (n=24)
  - POS tags + Morphosyntactic Descriptions
  - Linguistic diversity (synthesis + affixation)
- Word → Tag: Bidirectional LSTM + MLP
  - (Not analyzed)
  - **No word embeddings**



Model & Data

- Universal Dependencies (n=24)
  - POS tags + Morphosyntactic Descriptions
  - Linguistic diversity (synthesis + affixation)
- Word → Tag: Bidirectional LSTM + MLP
  - (Not analyzed)
  - **No word embeddings**
- Char → Word: Bidirectional LSTM
  - Char embedding size: 256



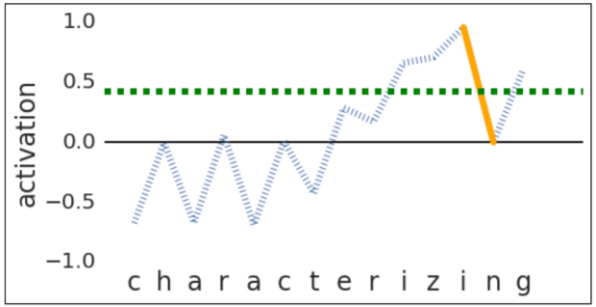
Analysis Metrics

Analysis Metrics

- Run model on training data words

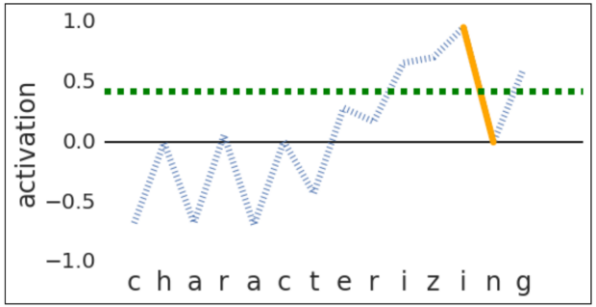
Analysis Metrics

- Run model on training data words
- Collect activation levels for each unit



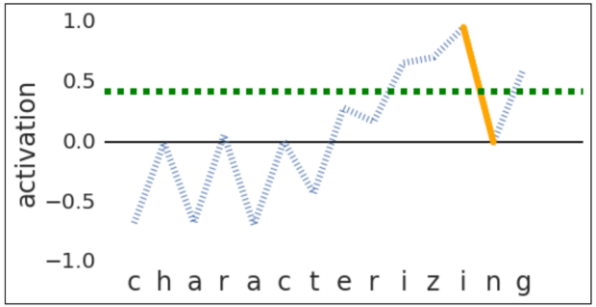
Analysis Metrics

- Run model on training data words
- Collect activation levels for each unit
- Aggregate to single measure
  (e.g. **average absolute** or **max-delta**)



Analysis Metrics

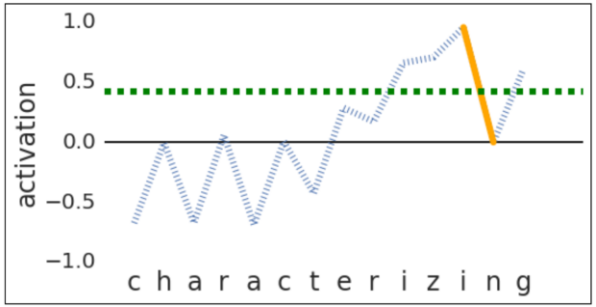
- Run model on training data words
- Collect activation levels for each unit
- Aggregate to single measure
  (e.g. **average absolute** or **max-delta**)
- Bin per unit over parts of speech



Unit 42	[0.0,0.1)	[0.1,0.2)	…	[0.9,1.0)
NOUN	8	2	…	40
VERB	20	0	…	4
…	…	…	…	…
ADJ	10	10	…	10

Analysis Metrics

- Run model on training data words
- Collect activation levels for each unit
- Aggregate to single measure (e.g. **average absolute** or **max-delta**)
- Bin per unit over parts of speech
- Mutual Information metric – POS Discrimination Index, or **PDI**
  - (Higher PDI = better discriminator)

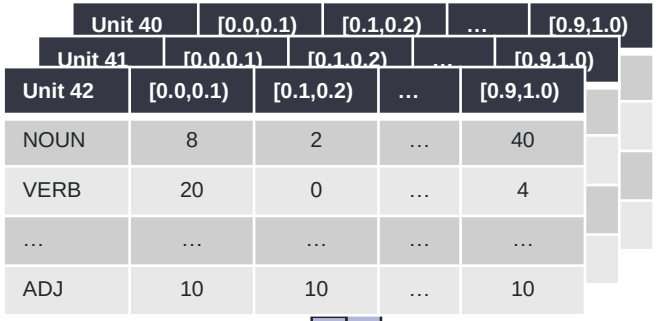


Unit 42	[0.0,0.1)	[0.1,0.2)	…	[0.9,1.0)
NOUN	8	2	…	40
VERB	20	0	…	4
…	…	…	…	…
ADJ	10	10	…	10

$$\sum_{t=1}^{T}\sum_{b=1}^{B} P(t,b)[\ln P(t,b) - \ln P(t) - \ln P(b)]$$

Analysis Metrics

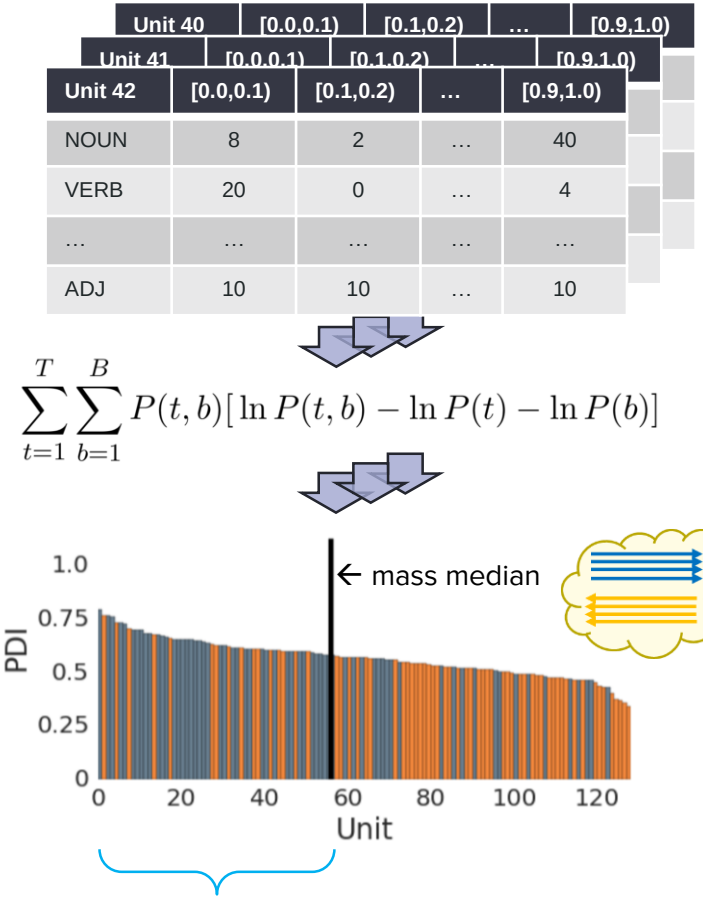
- Run model on training data words
- Collect activation levels for each unit
- Aggregate to single measure (e.g. **average absolute** or **max-delta**)
- Bin per unit over parts of speech
- Mutual Information metric – POS Discrimination Index, or **PDI**
  - (Higher PDI = better discriminator)
- Aggregate across units by



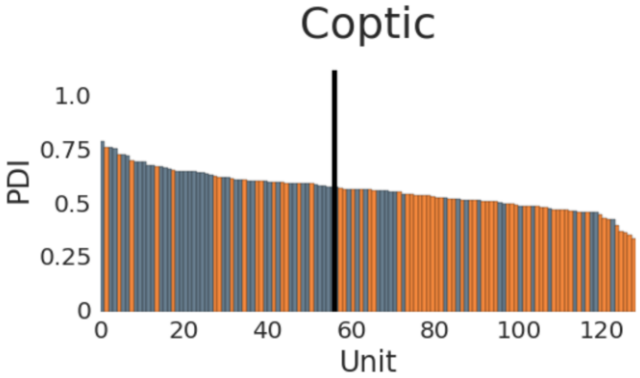
$$\sum_{t=1}^{T}\sum_{b=1}^{B} P(t,b)[\ln P(t,b) - \ln P(t) - \ln P(b)]$$

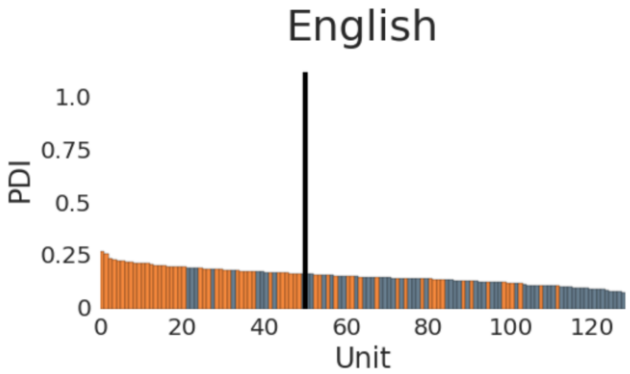
Analysis Metrics

- Run model on training data words
- Collect activation levels for each unit
- Aggregate to single measure (e.g. **average absolute** or **max-delta**)
- Bin per unit over parts of speech
- Mutual Information metric – POS Discrimination Index, or **PDI**
  - (Higher PDI = better discriminator)
- Aggregate across units by
  - Summing total mass
  - Reporting % of forward units before mass median



Findings (Cherry Pick)





- Coptic: agglutinative, prefixing
  - Large mass (easy to distinguish POS based on char sequence)
  - Forward-heavy (71%)

- English: fusional, suffixing
  - Small mass (hard to capture POS)
  - Backward-heavy (80%)

Findings (General Trends)

	Total PDI mass
Tamil	71.0
Irish	62.0
Coptic	58.1
Hungarian	47.9
Greek	31.2
Turkish	30.1
Russian	25.9
Thai	25.9
Ukrainian	25.0
Vietnamese	24.2
Chinese	23.8
Danish	21.7
Swedish	20.8
Basque	20.6
Indonesian	20.3
Latvian	17.0
Spanish	16.1
English	16.0
Bulgarian	15.6
Italian	14.1
Arabic	12.6
Hebrew	11.4
Persian	10.3
Hindi	8.4

Findings (General Trends)

- 4/5 agglutinatives hold 4/6 top total-mass positions

	Total PDI mass
Tamil	71.0
Irish	62.0
Coptic	58.1
Hungarian	47.9
Greek	31.2
Turkish	30.1
Russian	25.9
Thai	25.9
Ukrainian	25.0
Vietnamese	24.2
Chinese	23.8
Danish	21.7
Swedish	20.8
Basque	20.6
Indonesian	20.3
Latvian	17.0
Spanish	16.1
English	16.0
Bulgarian	15.6
Italian	14.1
Arabic	12.6
Hebrew	11.4
Persian	10.3
Hindi	8.4

Findings (General Trends)

- 4/5 agglutinatives hold 4/6 top total-mass positions

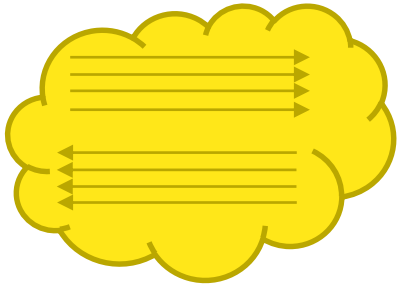
- 2/2 introflexives in bottom 2/4 spots (Persian and Hindi below, both fusional w/ non-Latin charsets)

	Total PDI mass
Tamil	71.0
Irish	62.0
Coptic	58.1
Hungarian	47.9
Greek	31.2
Turkish	30.1
Russian	25.9
Thai	25.9
Ukrainian	25.0
Vietnamese	24.2
Chinese	23.8
Danish	21.7
Swedish	20.8
Basque	20.6
Indonesian	20.3
Latvian	17.0
Spanish	16.1
English	16.0
Bulgarian	15.6
Italian	14.1
Arabic	12.6
Hebrew	11.4
Persian	10.3
Hindi	8.4

Direction Balance Study

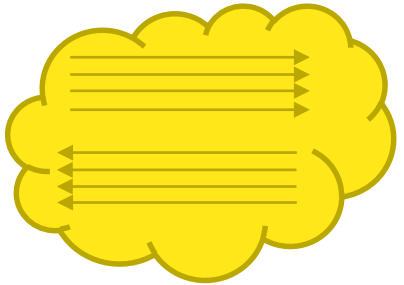
Direction Balance Study

- Some languages might not need two equal LSTM directions

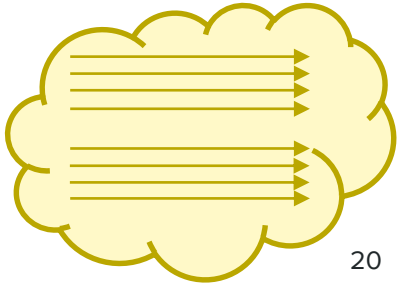


Direction Balance Study

- Some languages might not need two equal LSTM directions

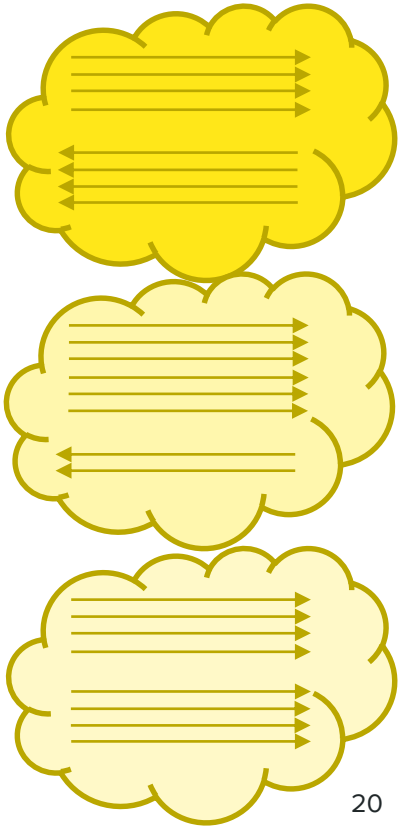


- What if… they don’t need one of them at all?



Direction Balance Study

- Some languages might not need two equal LSTM directions
- What if they need them in a different balance? Somewhere in the middle?
- What if... they don't need one of them at all?



Balance Study – Results

Balance Study – Results

- Can unidirectional models outperform bidirectionals?

Balance Study – Results

- Can unidirectional models outperform bidirectionals?
- **Yes**.
  - Especially on agglutinative languages and on suffixing languages
  - Fully-forward better than fully-backward

Balance Study – Results

- Can unidirectional models outperform bidirectionals?
- **Yes**.
  - Especially on agglutinative languages and on suffixing languages
  - Fully-forward better than fully-backward
  - MAJOR caveat – 128x128 > 2*(64x64)

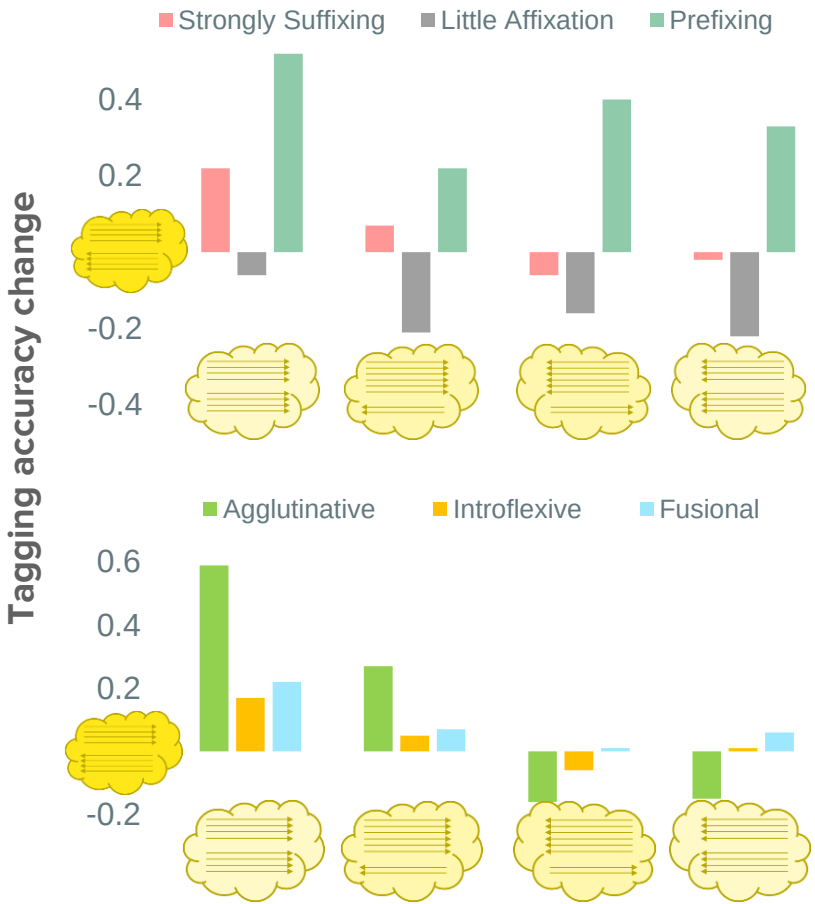
Balance Study – Results

- Can unidirectional models outperform bidirectionals?
- **Yes**.
  - Especially on agglutinative languages and on suffixing languages
  - Fully-forward better than fully-backward
  - MAJOR caveat – 128x128 > 2*(64x64)
- Is there a sweet spot in the middle?

Balance Study – Results

- Can unidirectional models outperform bidirectionals?
- **Yes**.
  - Especially on agglutinative languages and on suffixing languages
  - Fully-forward better than fully-backward
  - MAJOR caveat – 128x128 > 2*(64x64)

- Is there a sweet spot in the middle?
- **Not that we can tell**.



Summary + Open Ends

Summary + Open Ends

- Introduced **PDI** to aggregate information from hidden units, some applicability to language characterization

Summary + Open Ends

- Introduced **PDI** to aggregate information from hidden units, some applicability to language characterization
  - Extensible to any <instance, unit> metric on any neural classifier

Summary + Open Ends

- Introduced **PDI** to aggregate information from hidden units, some applicability to language characterization
  - Extensible to any <instance, unit> metric on any neural classifier
- Found substantial differences between differently-balanced recurrent models

Summary + Open Ends

- Introduced **PDI** to aggregate information from hidden units, some applicability to language characterization
  - Extensible to any <instance, unit> metric on any neural classifier
- Found substantial differences between differently-balanced recurrent models

- Are we quantifying **data** instead of **languages**?

Summary + Open Ends

- Introduced **PDI** to aggregate information from hidden units, some applicability to language characterization
  - Extensible to any <instance, unit> metric on any neural classifier
- Found substantial differences between differently-balanced recurrent models
- Are we quantifying **data** instead of **languages**?
- Affixing: many languages (e.g. English) have higher PDI for **backward** units, but fare better with more **forward** units. Is this:

Summary + Open Ends

- Introduced **PDI** to aggregate information from hidden units, some applicability to language characterization
  - Extensible to any <instance, unit> metric on any neural classifier
- Found substantial differences between differently-balanced recurrent models
- Are we quantifying **data** instead of **languages**?
- Affixing: many languages (e.g. English) have higher PDI for **backward** units, but fare better with more **forward** units. Is this:
  - A saturation effect?

Summary + Open Ends

- Introduced **PDI** to aggregate information from hidden units, some applicability to language characterization
  - Extensible to any <instance, unit> metric on any neural classifier
- Found substantial differences between differently-balanced recurrent models
- Are we quantifying **data** instead of **languages**?
- Affixing: many languages (e.g. English) have higher PDI for **backward** units, but fare better with more **forward** units. Is this:
  - A saturation effect?
  - Fault in assuming PDI measures unit importance?

Thank You!

https://github.com/ruyimarone/character-eyes

uvp@gatech.edu
mmarone6@gatech.edu