A Distributional and Orthographic Aggregation Model for English Derivational Morphology

Daniel Deutsch,* John Hewitt,* and Dan Roth

*equal contribution

Co-Authors

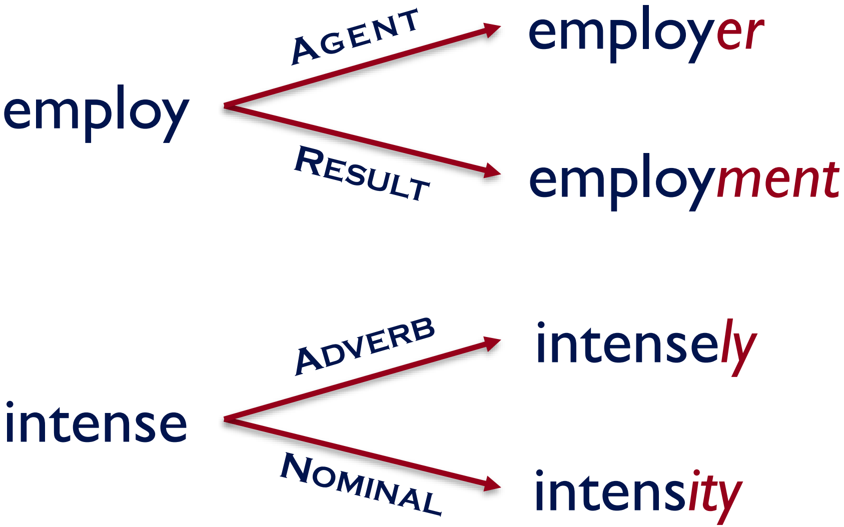
John Hewitt

Co-First Author

Dan Roth

Advisor

Derivational Morphology



Derivational Morphology

employ $\xrightarrow{\text{Agent}}$ employ*er*

employ $\xrightarrow{\text{Result}}$ employ*ment*

intense $\xrightarrow{\text{Adverb}}$ intense*ly*

intense $\xrightarrow{\text{Nominal}}$ intens*ity*

root word $\xrightarrow{\text{TRANSFORMATION}}$ derived word

Derivational Morphology

employ $\xrightarrow{\text{AGENT}}$ employ*er*

employ $\xrightarrow{\text{RESULT}}$ employ*ment*

intense $\xrightarrow{\text{ADVERB}}$ intense*ly*

intense $\xrightarrow{\text{NOMINAL}}$ intens*ity*

root word $\xrightarrow{\text{TRANSFORMATION}}$ derived word

Motivation

- Machine translation
- Text simplification
- Language generation

Challenges

- Suffix ambiguity
- Orthographic irregularity

Suffix Ambiguity

"I have an observa*ment*!"

Suffix Ambiguity

"I have an observa*ment*!"

ground

- ground*ing*
- *ground*ation*
- *ground*ment*
- *ground*al*

RESULT

Suffix Ambiguity

"I have an observa*ment*!"

ground → grounding, *groundation, *groundment, *groundal

RESULT

valid → validity, *validness

NOMINAL

Orthographic Irregularity

speak $\xrightarrow{\text{RESULT}}$ spee*ch*

Orthographic Irregularity

speak $\xrightarrow{\text{RESULT}}$ spee*ch*

creak $\xrightarrow{\text{RESULT}}$

Orthographic Irregularity

speak $\xrightarrow{\text{RESULT}}$ spee*ch*

creak $\xrightarrow{\text{RESULT}}$ *cree*ch*

Orthographic Irregularity

speak $\xrightarrow{\text{RESULT}}$ speech

creak $\xrightarrow{\text{RESULT}}$ *~~creech~~ creaking

Orthographic Irregularity

speak $\xrightarrow{\text{RESULT}}$ spee*ch*

creak $\xrightarrow{\text{RESULT}}$ *~~creech~~ creak*ing*

erupt $\xrightarrow{\text{RESULT}}$ erupt*ion*

Penn Engineering

Orthographic Irregularity

speak $\xrightarrow{\textsc{Result}}$ spee*ch*

creak $\xrightarrow{\textsc{Result}}$ *~~cree*ch*~~ creak*ing*

erupt $\xrightarrow{\textsc{Result}}$ erupt*ion*

bankrupt $\xrightarrow{\textsc{Result}}$

Orthographic Irregularity

speak $\xrightarrow{\text{RESULT}}$ speech

creak $\xrightarrow{\text{RESULT}}$ *~~creech~~ creaking

erupt $\xrightarrow{\text{RESULT}}$ eruption

bankrupt $\xrightarrow{\text{RESULT}}$ *bankruption

Orthographic Irregularity

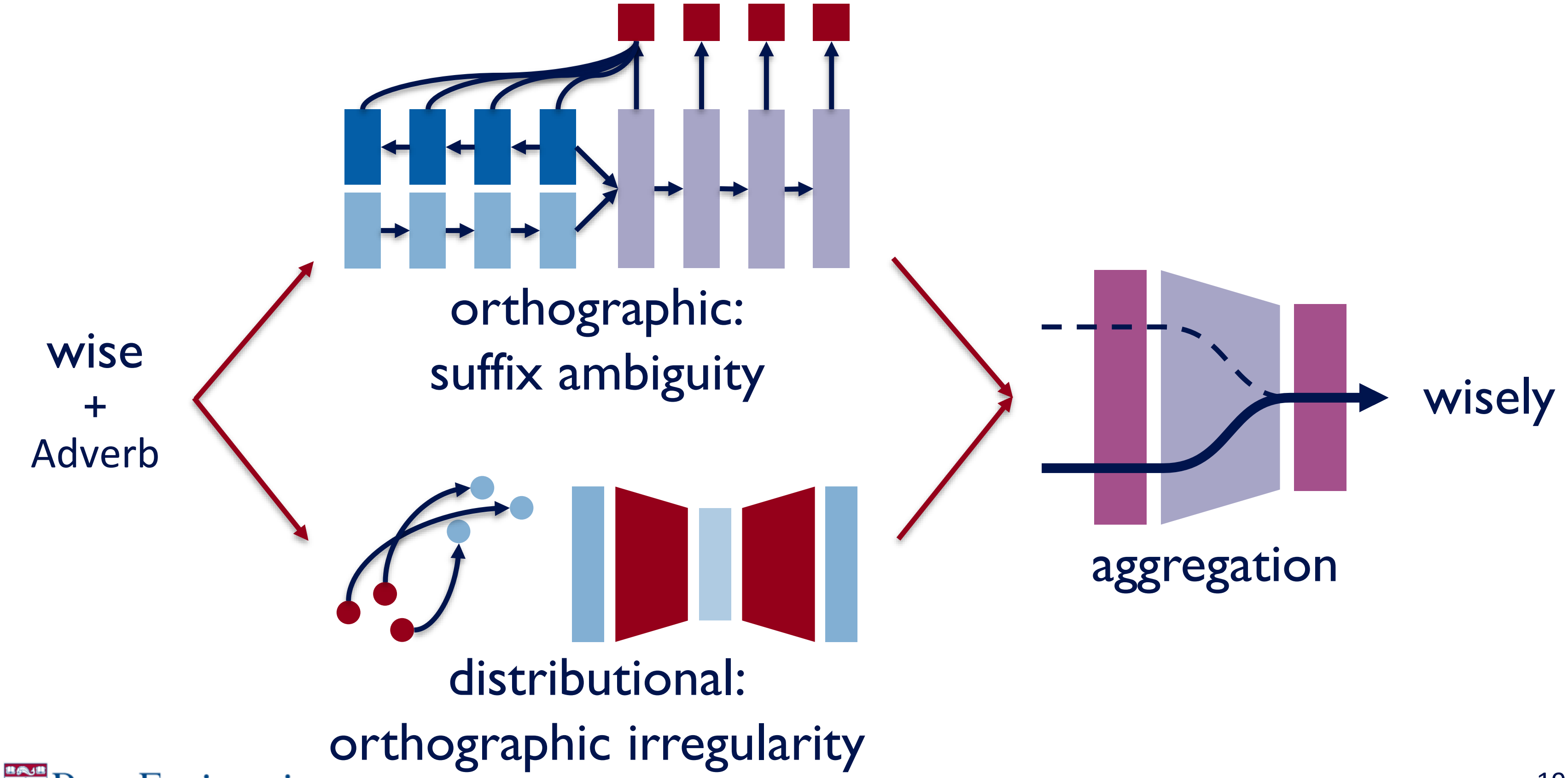
speak $\xrightarrow{\textsc{Result}}$ spee*ch*

creak $\xrightarrow{\textsc{Result}}$ *~~cree*ch*~~ creak*ing*

erupt $\xrightarrow{\textsc{Result}}$ erupt*ion*

bankrupt $\xrightarrow{\textsc{Result}}$ *~~bankrupt*ion*~~ bankrupt*cy*

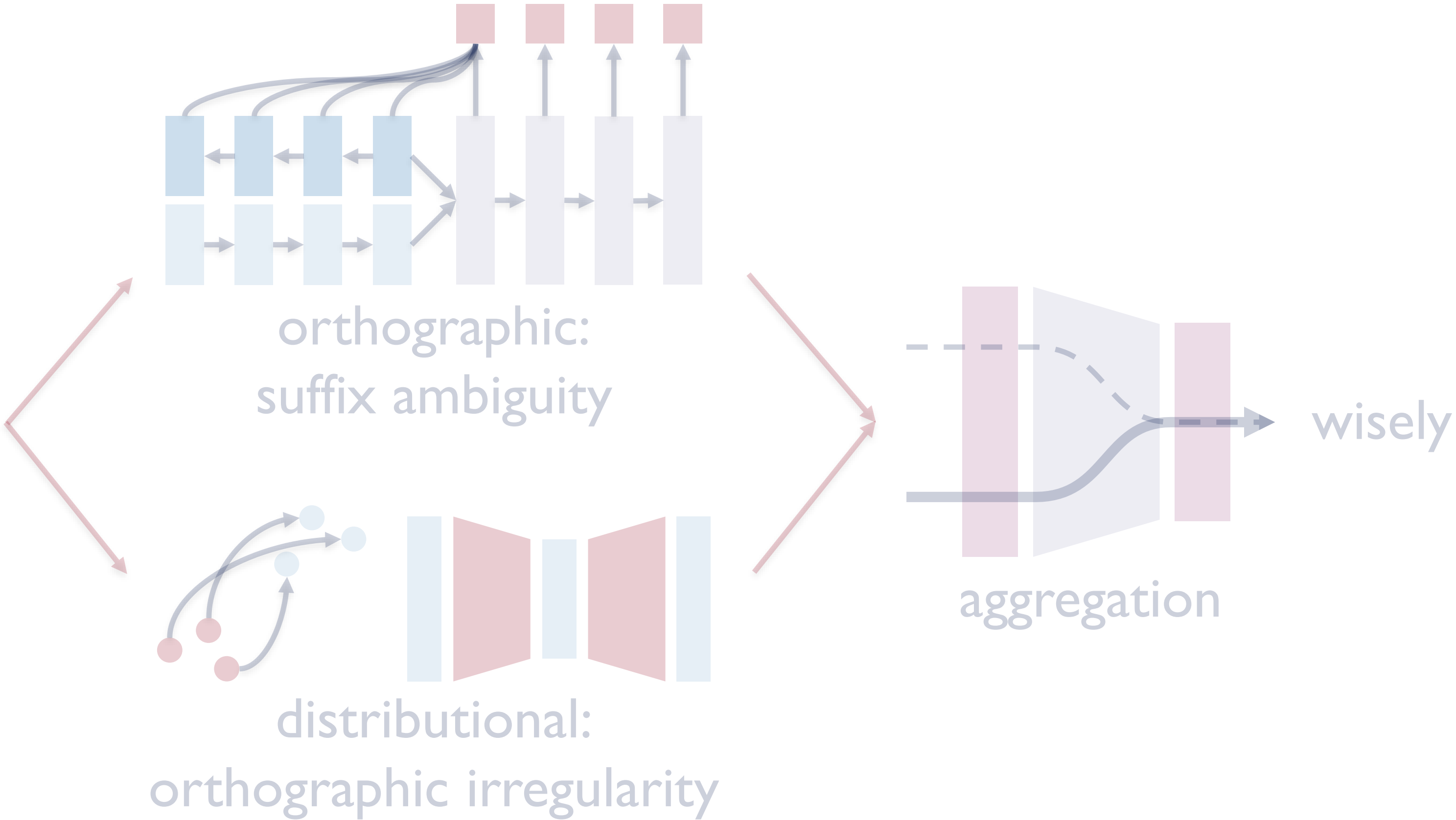
Model Overview



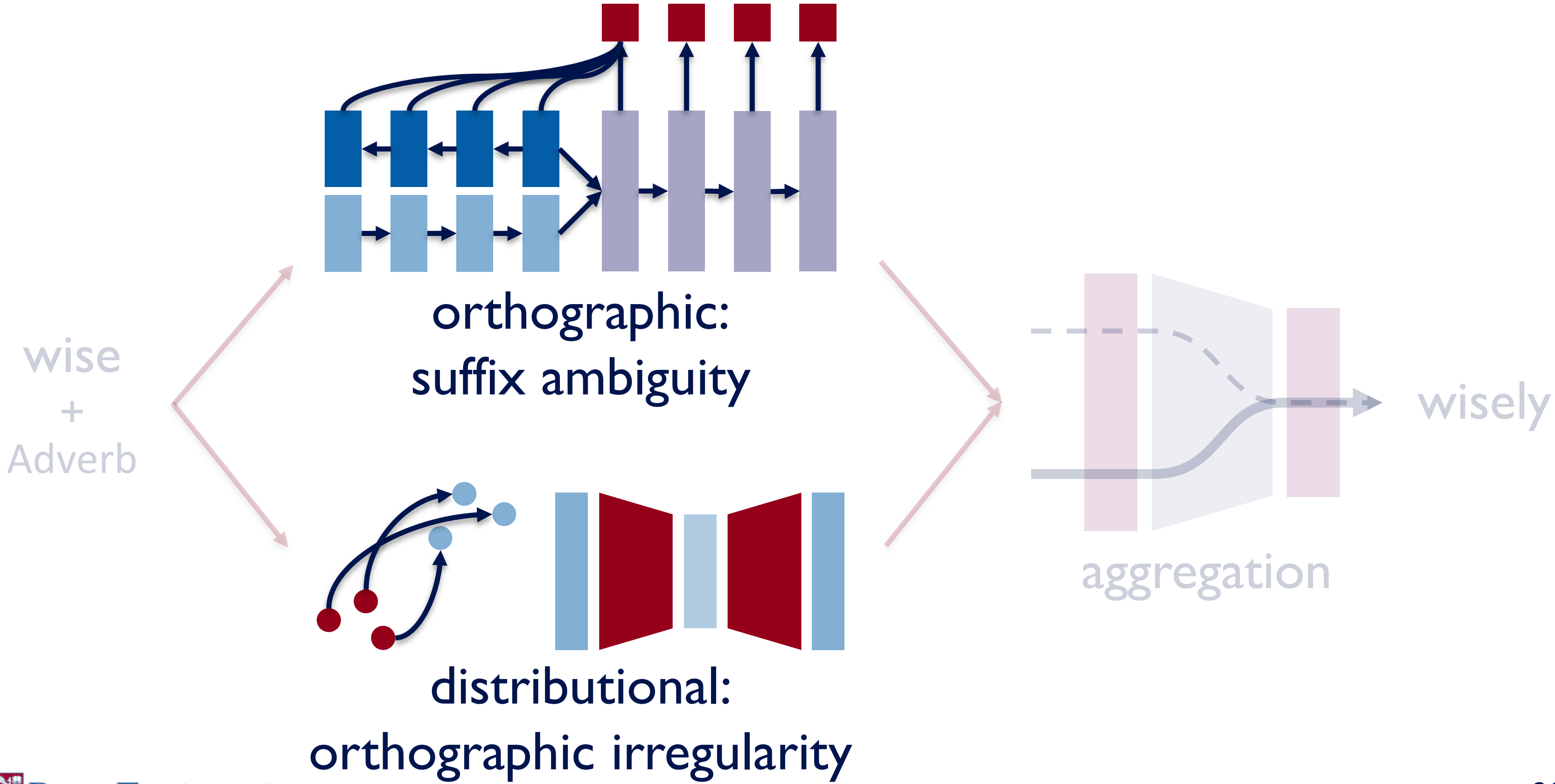
Penn Engineering

Model Overview

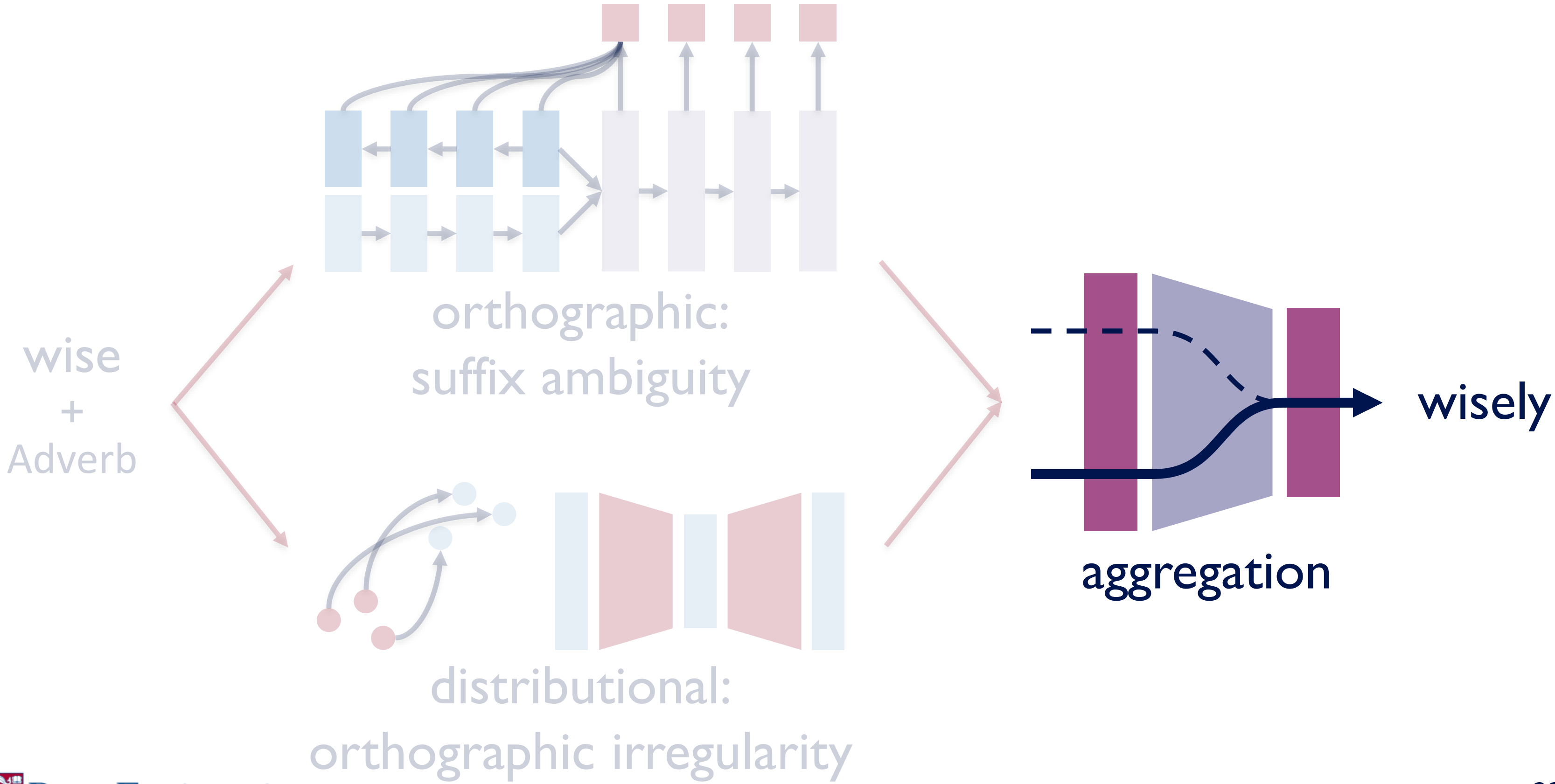
wise
+
Adverb



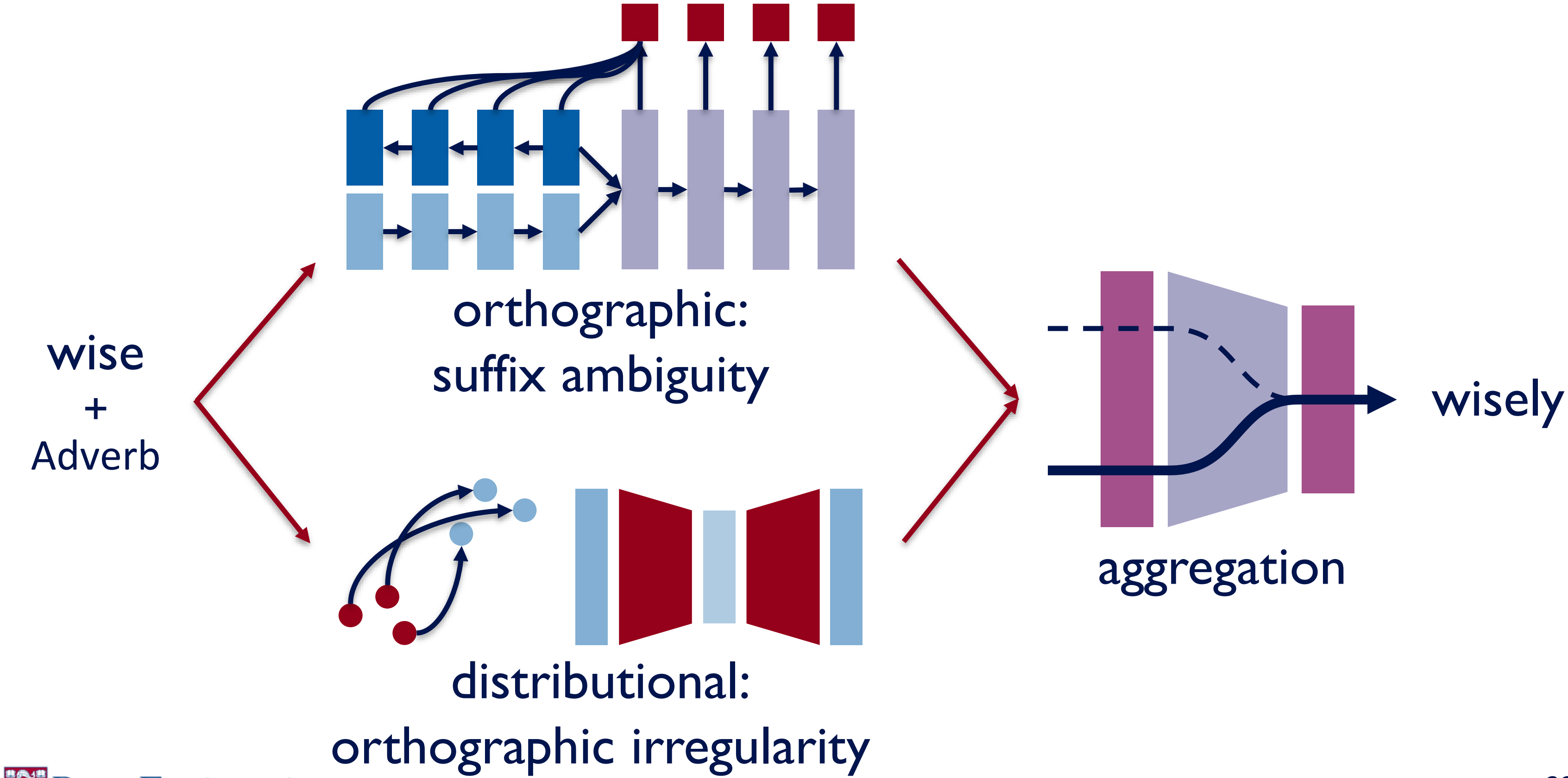
Model Overview



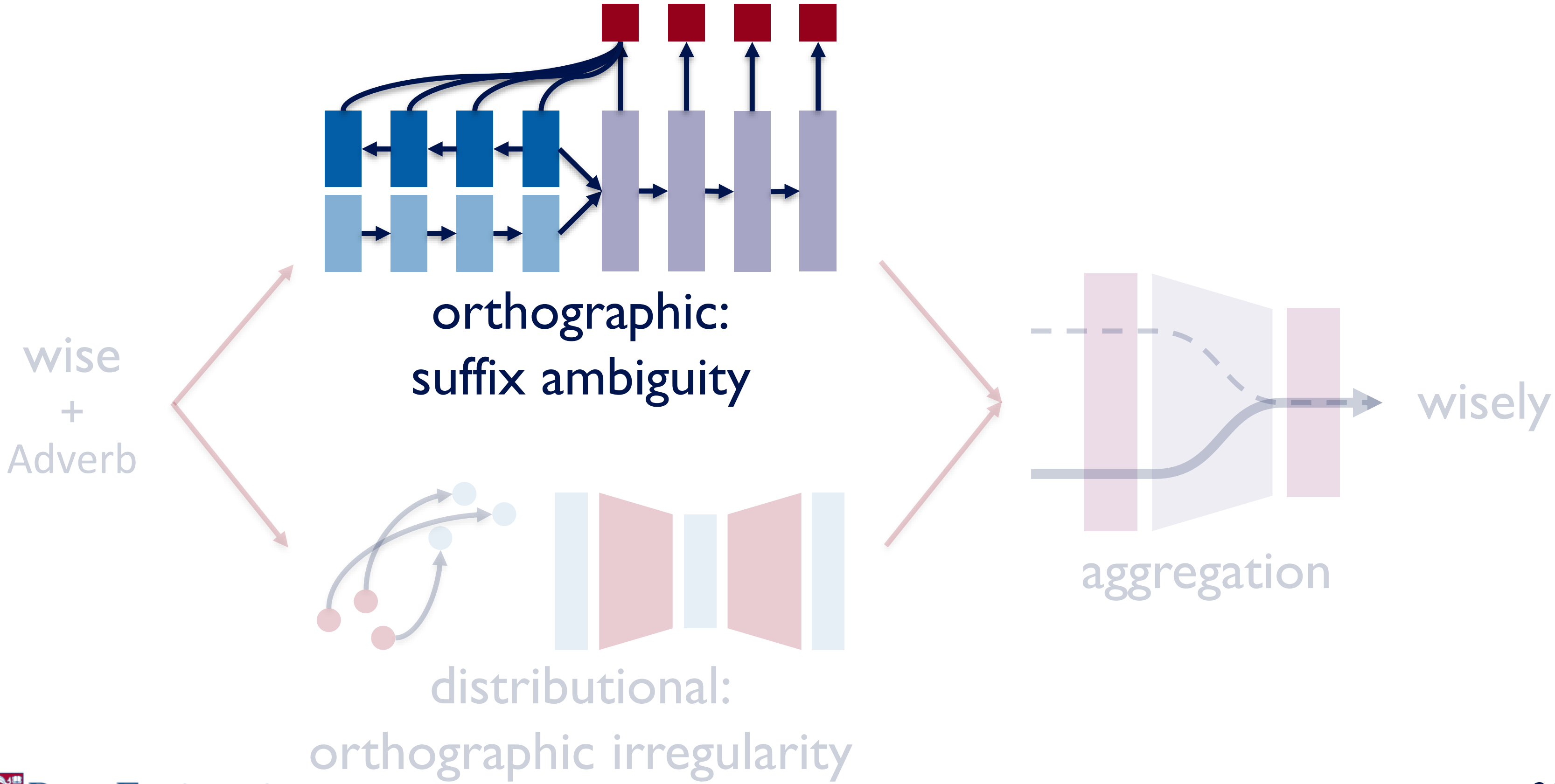
Model Overview



Model Overview



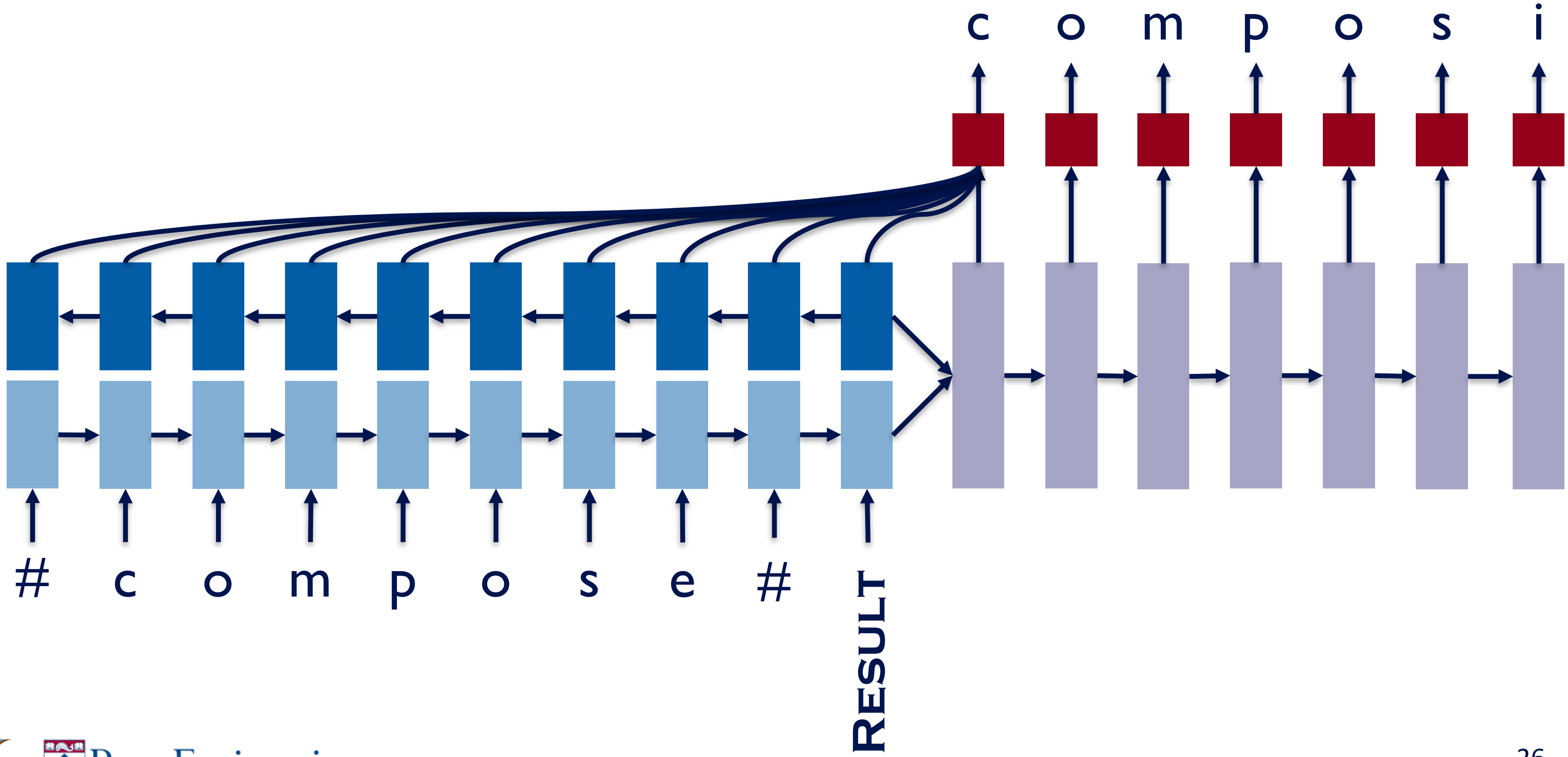
Model Overview



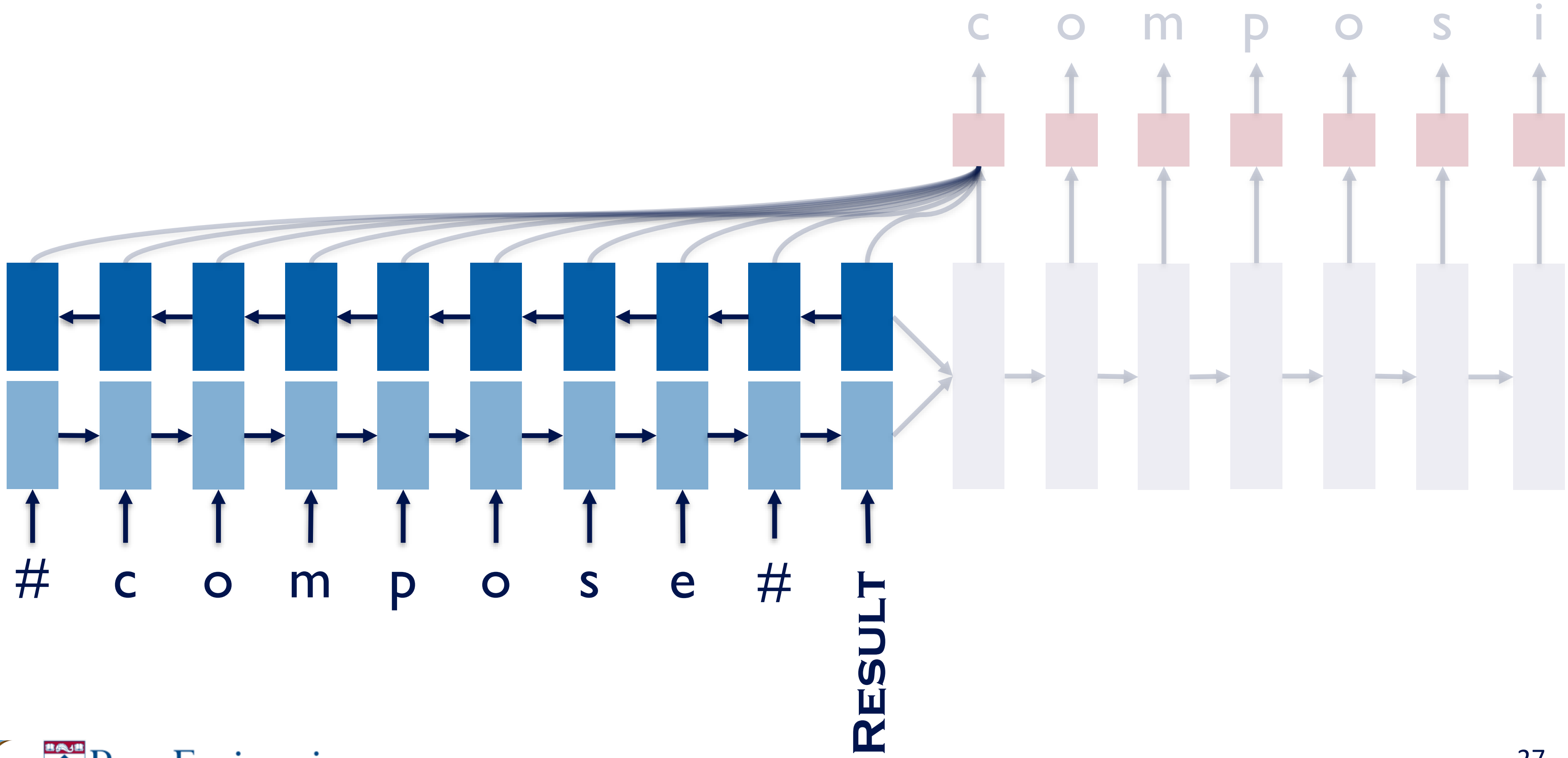
Orthographic Model

- Seq2Seq baseline
- Dictionary-constrained decoding
- Reranking with frequency information

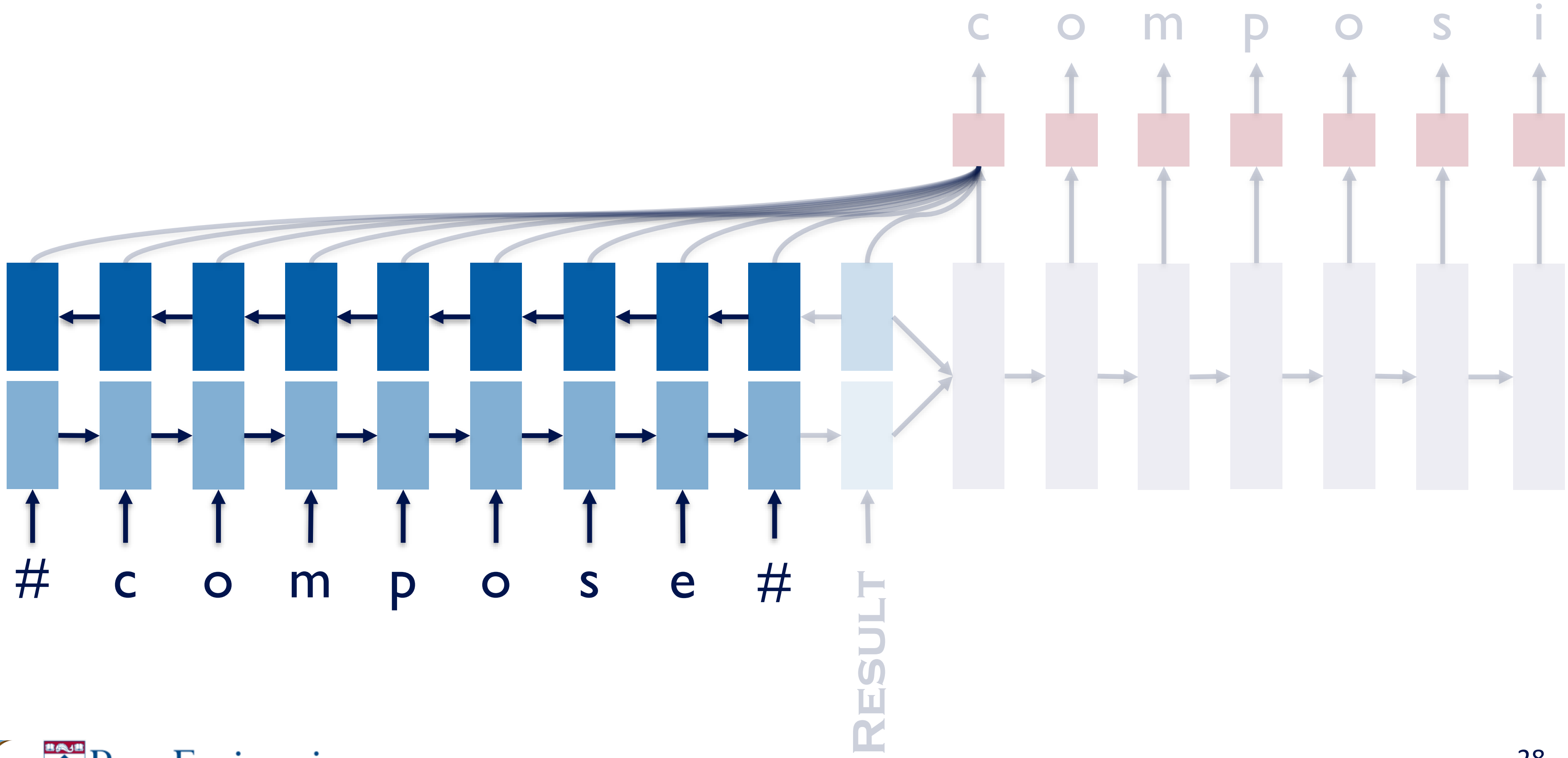
Seq2Seq Baseline



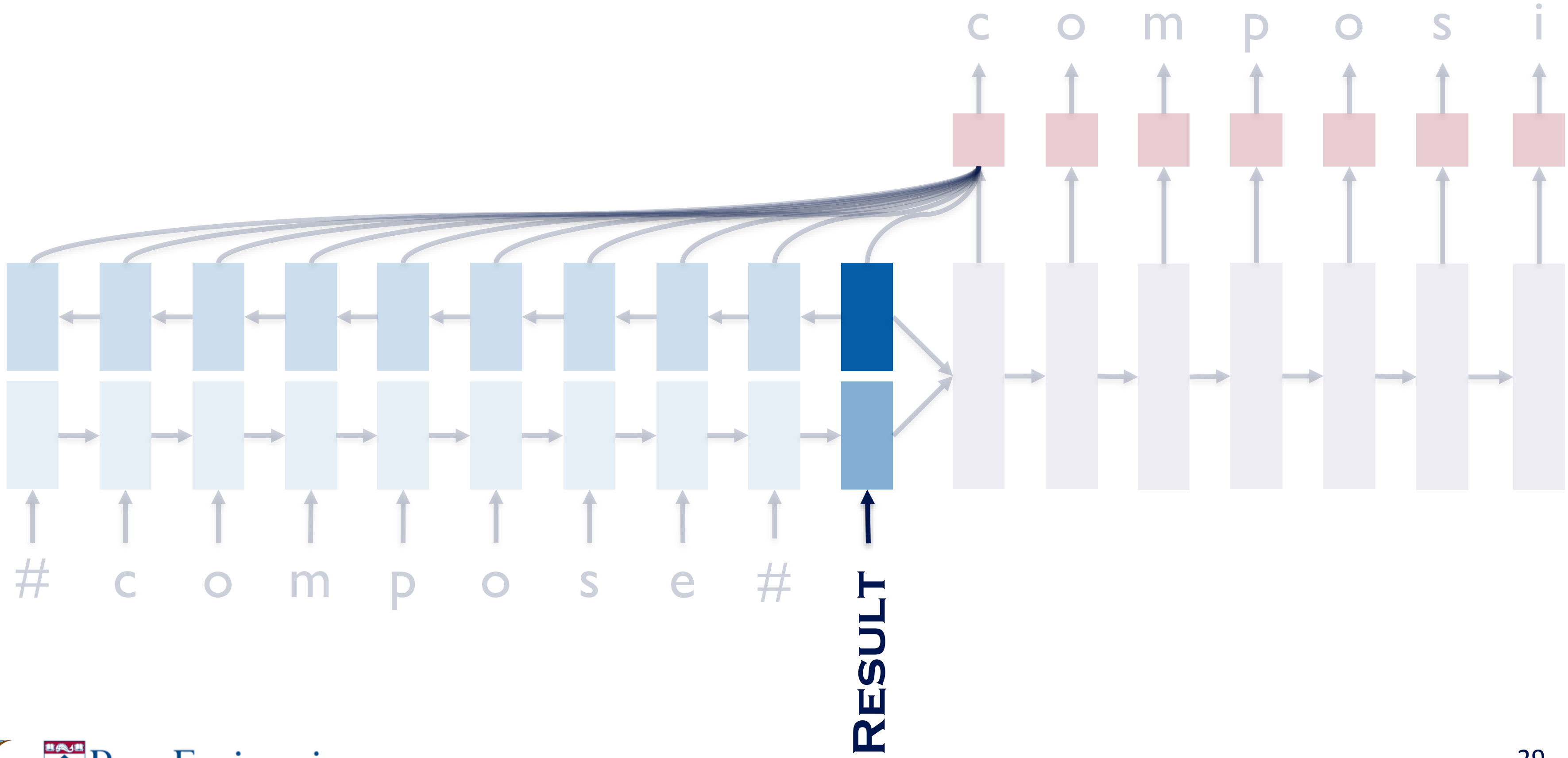
Seq2Seq Baseline



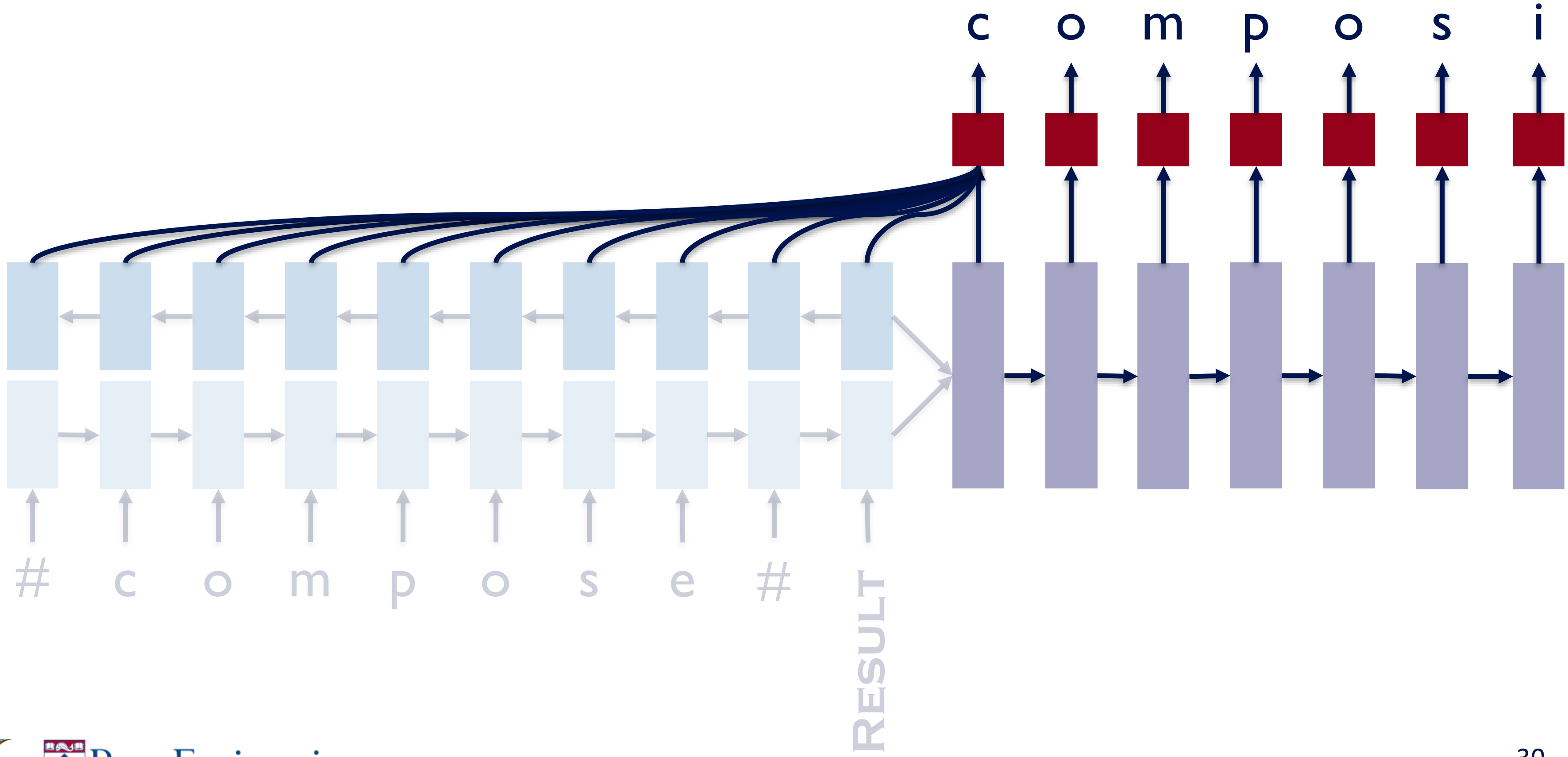
Seq2Seq Baseline



Seq2Seq Baseline



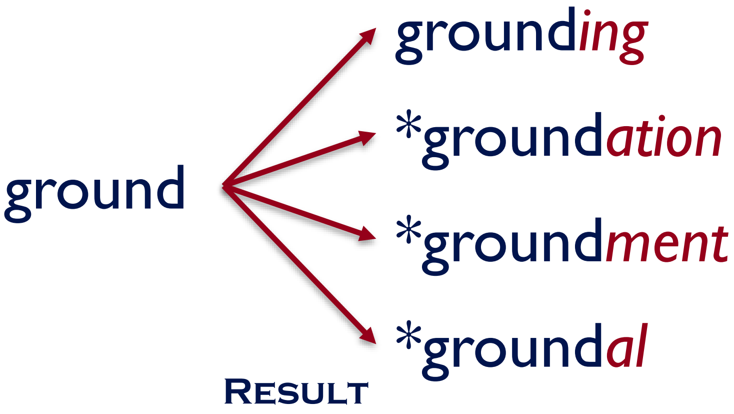
Seq2Seq Baseline



Dictionary-Constrained Decoding

- Seq2Seq models generate many unattested words, but are reasonable guesses

Suffix Ambiguity

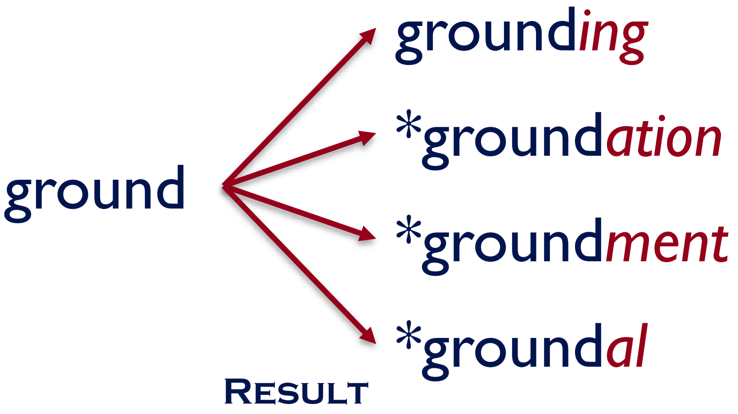


Penn Engineering

Dictionary-Constrained Decoding

- Seq2Seq models generate many unattested words, but are reasonable guesses
- Intuition: constrain model to only generate known words

Suffix Ambiguity



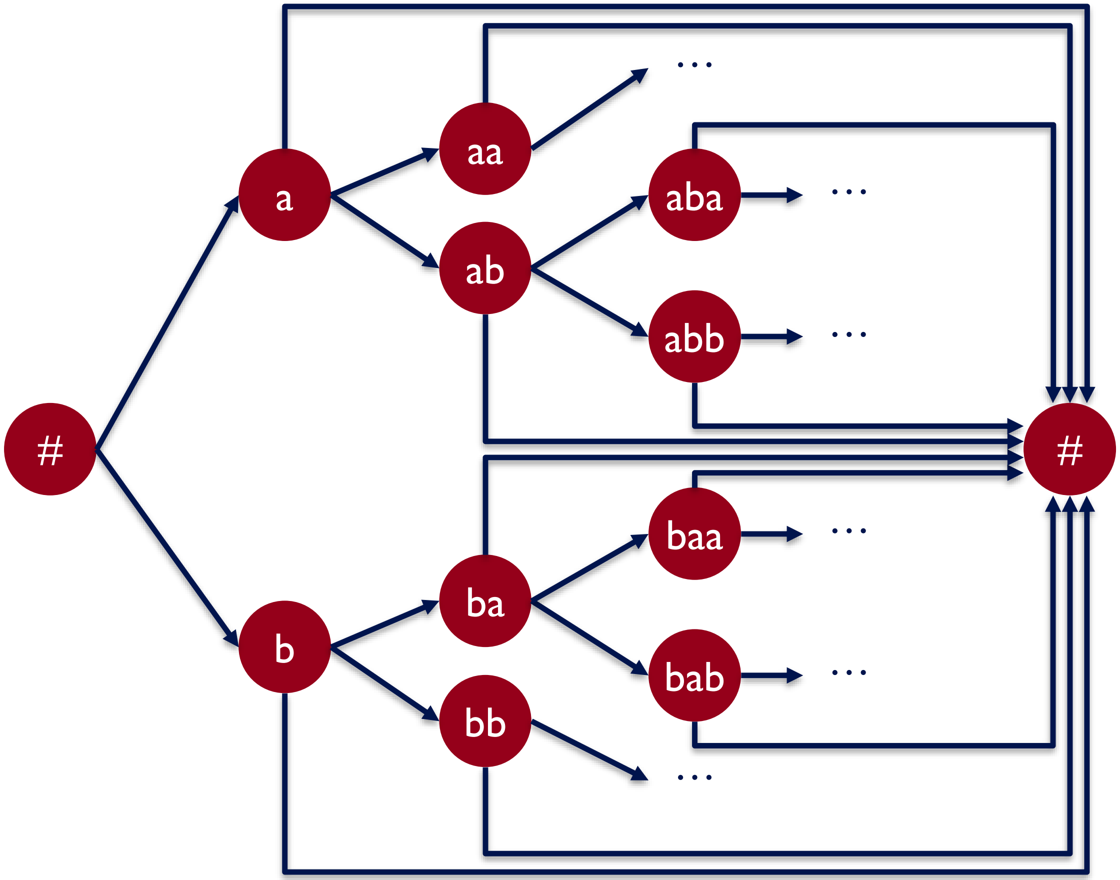
Penn Engineering

Dictionary-Constrained Decoding

$$\Sigma = \{a, b, \#\}$$

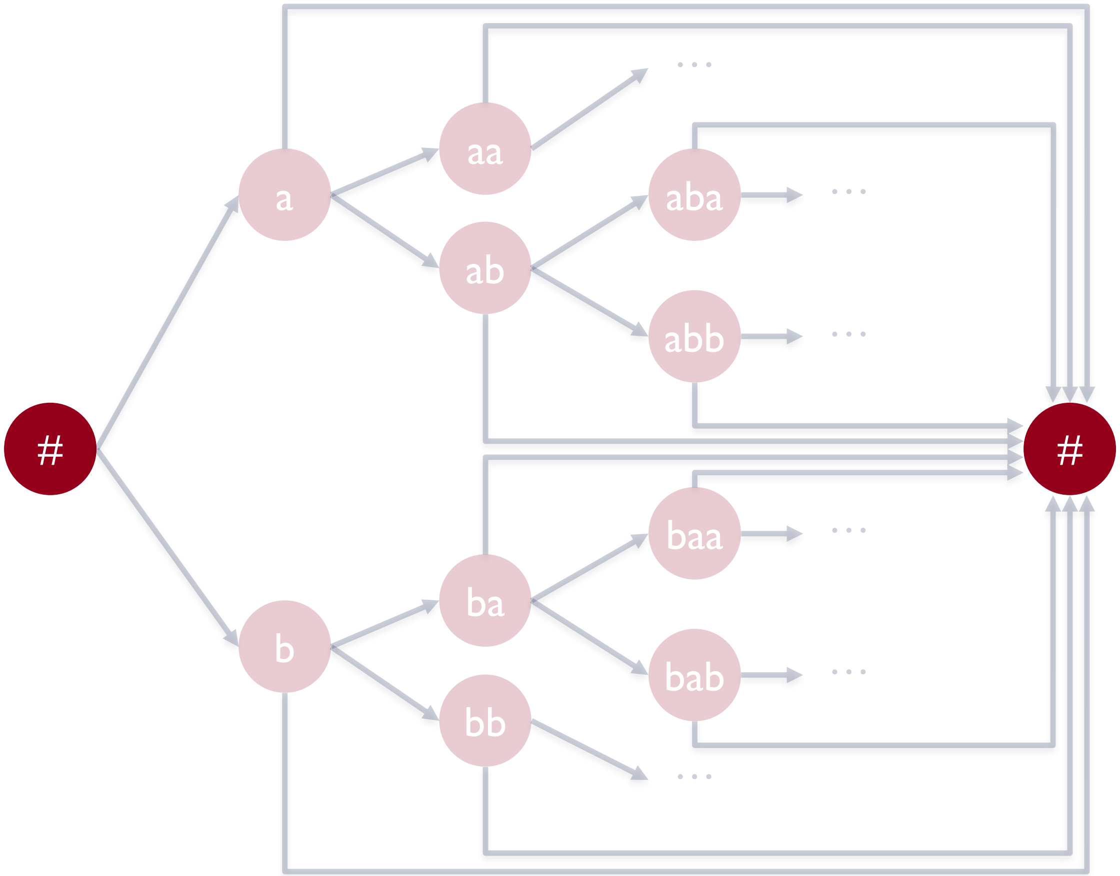
Dictionary-Constrained Decoding

$\Sigma = \{a, b, \#\}$



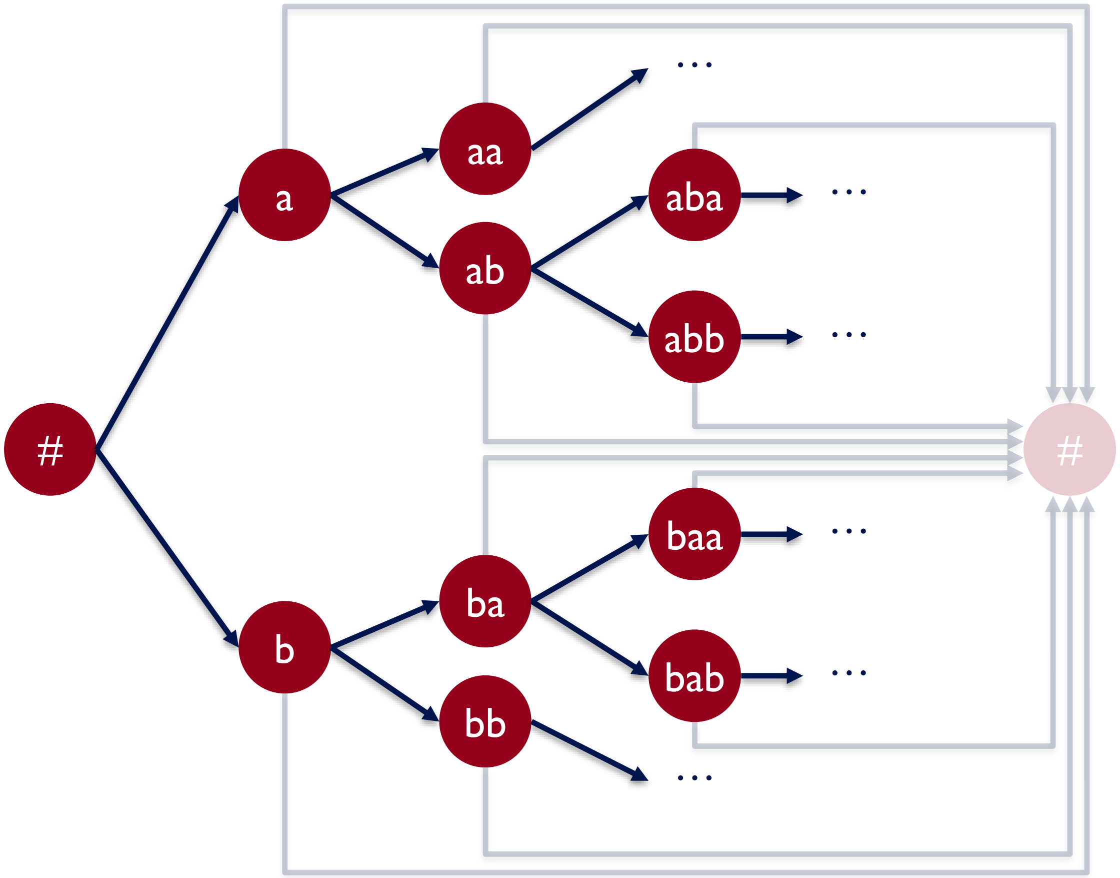
Dictionary-Constrained Decoding

$$\Sigma = \{a, b, \#\}$$



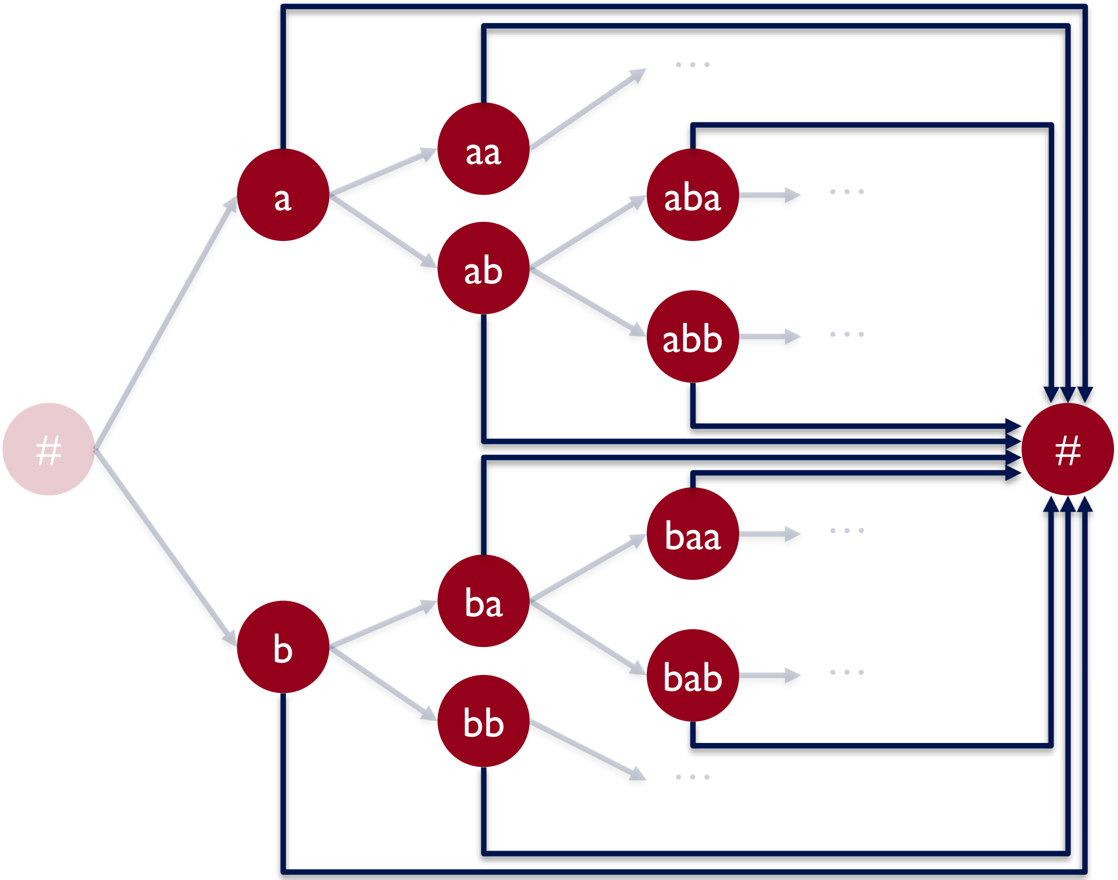
Dictionary-Constrained Decoding

$$\Sigma = \{a, b, \#\}$$



Dictionary-Constrained Decoding

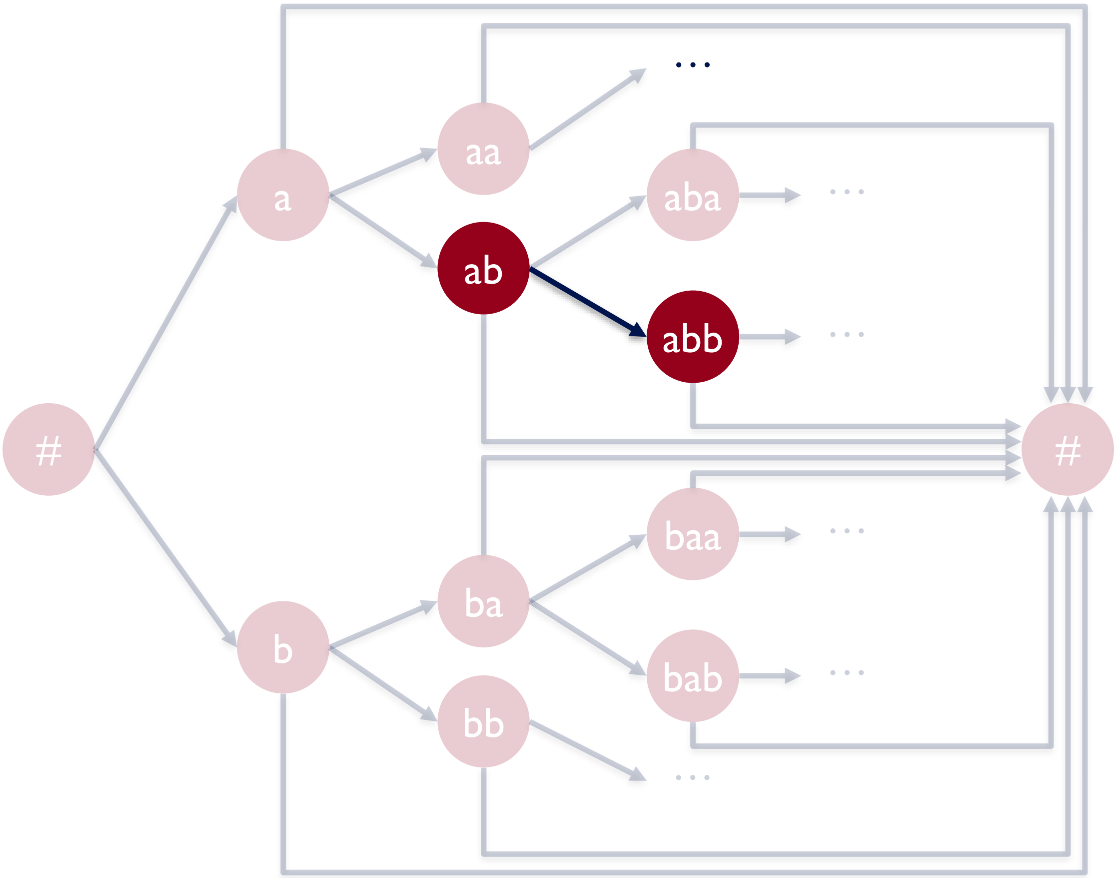
$$\Sigma = \{a, b, \#\}$$



Dictionary-Constrained Decoding

$$\Sigma = \{a, b, \#\}$$

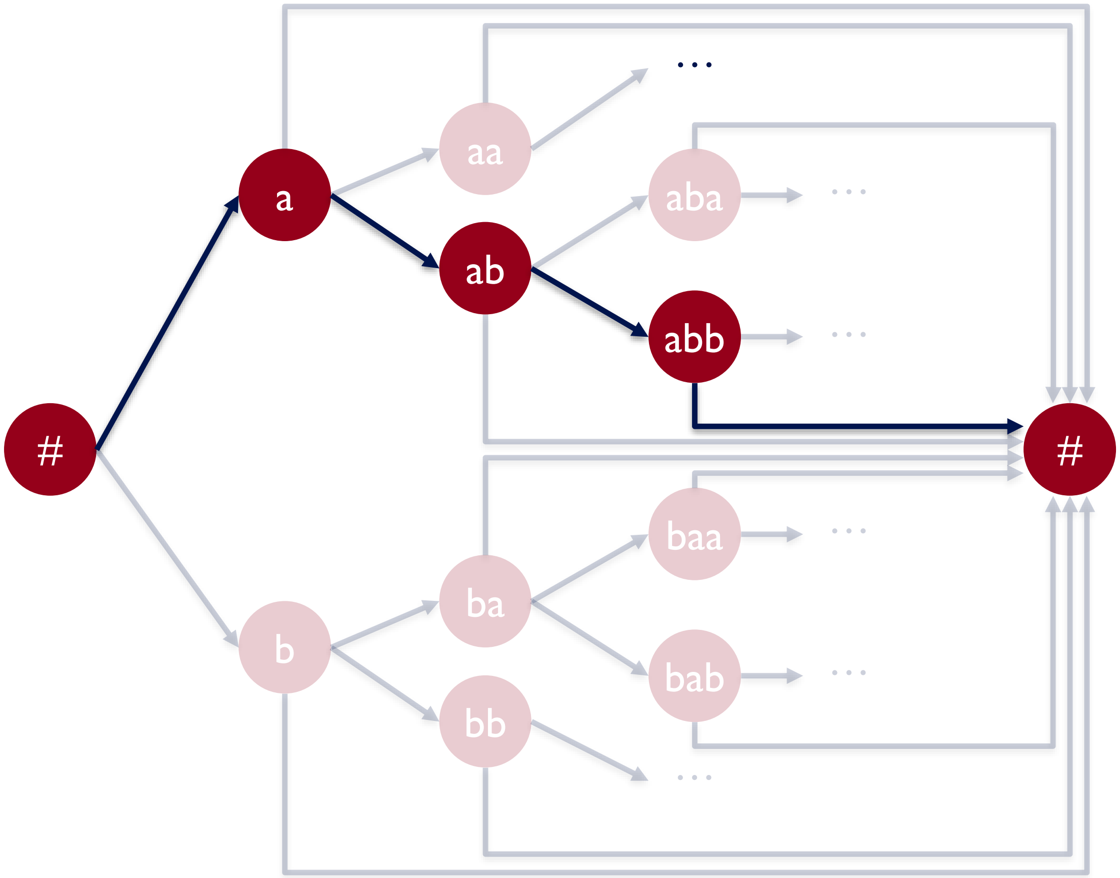
$$-\log p_\theta(b \mid ab)$$



Dictionary-Constrained Decoding

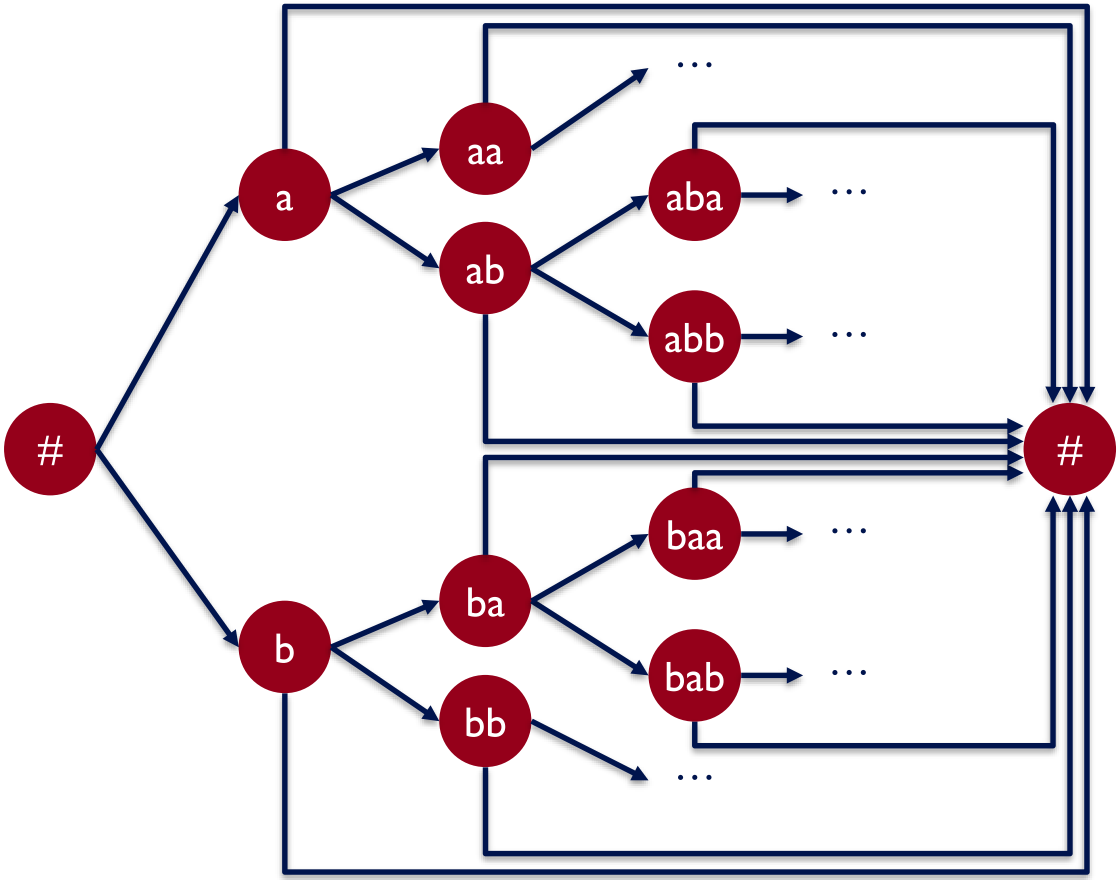
$$\Sigma = \{a, b, \#\}$$

$$-\log p_\theta(abb)$$



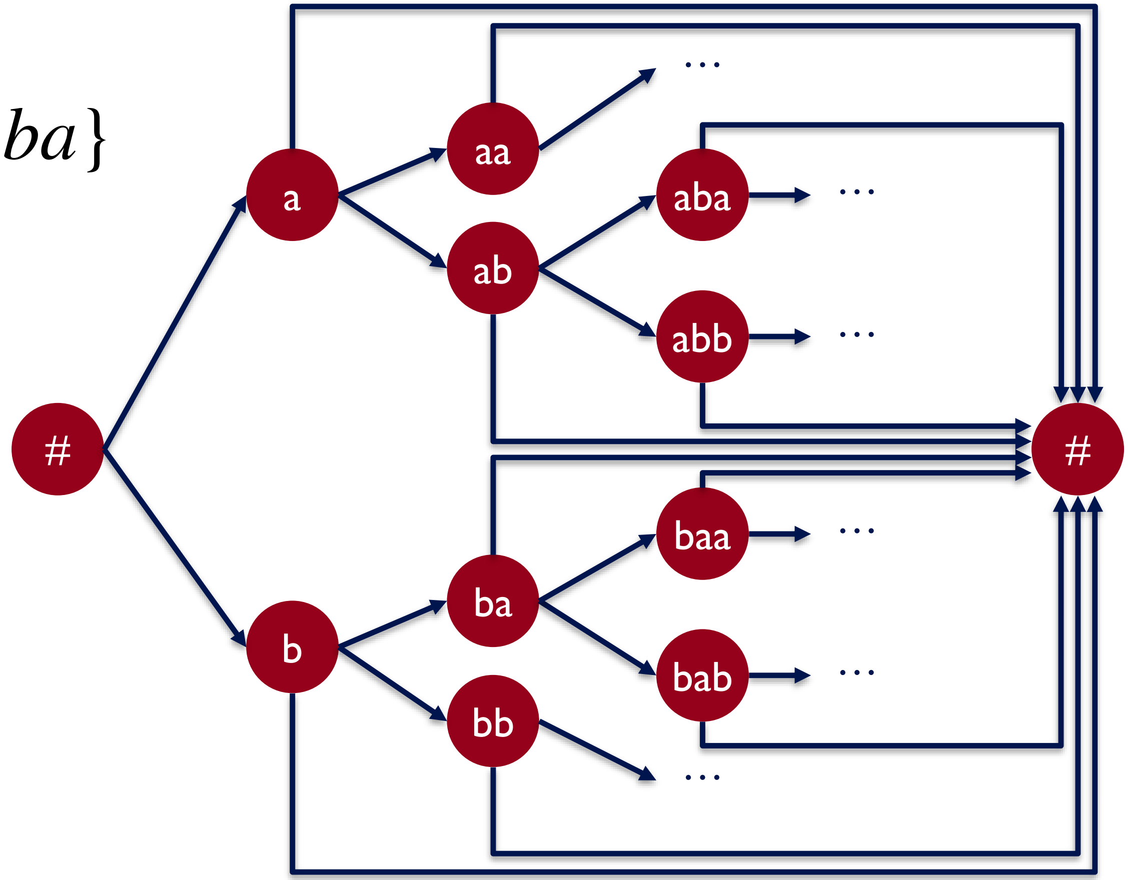
Dictionary-Constrained Decoding

$$\Sigma = \{a, b, \#\}$$



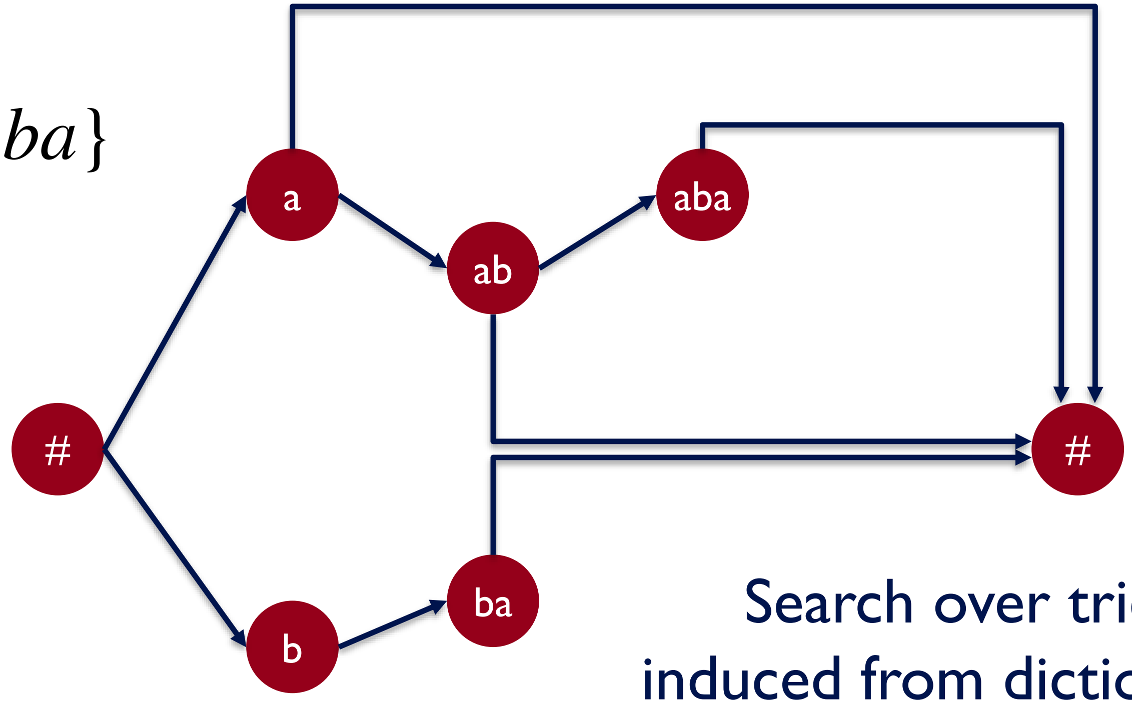
Dictionary-Constrained Decoding

$$\Sigma = \{a, b, \#\}$$
$$L = \{a, ab, aba, ba\}$$



Dictionary-Constrained Decoding

$$\Sigma = \{a, b, \#\}$$
$$L = \{a, ab, aba, ba\}$$



Search over trie induced from dictionary

Reranking with Frequency Information

refute $\xrightarrow{\text{RESULT}}$

Reranking with Frequency Information

refute $\xrightarrow{\text{RESULT}}$

Model Output	Model Score
refution	-1.1
refutation	-1.2
refut	-4.8
refuty	-5.6
refutat	-8.7

Reranking with Frequency Information

refute $\xrightarrow{\text{RESULT}}$

Model Output	Model Score
refution	-1.1
refutation	-1.2
refut	-4.8
refuty	-5.6
refutat	-8.7

Penn Engineering

Reranking with Frequency Information

refute $\xrightarrow{\text{RESULT}}$

Model Output	Model Score	Log Corpus Freq
refution	-1.1	5.0
refutation	-1.2	14.3
refut	-4.8	7.4
refuty	-5.6	0.1
refutat	-8.7	8.6

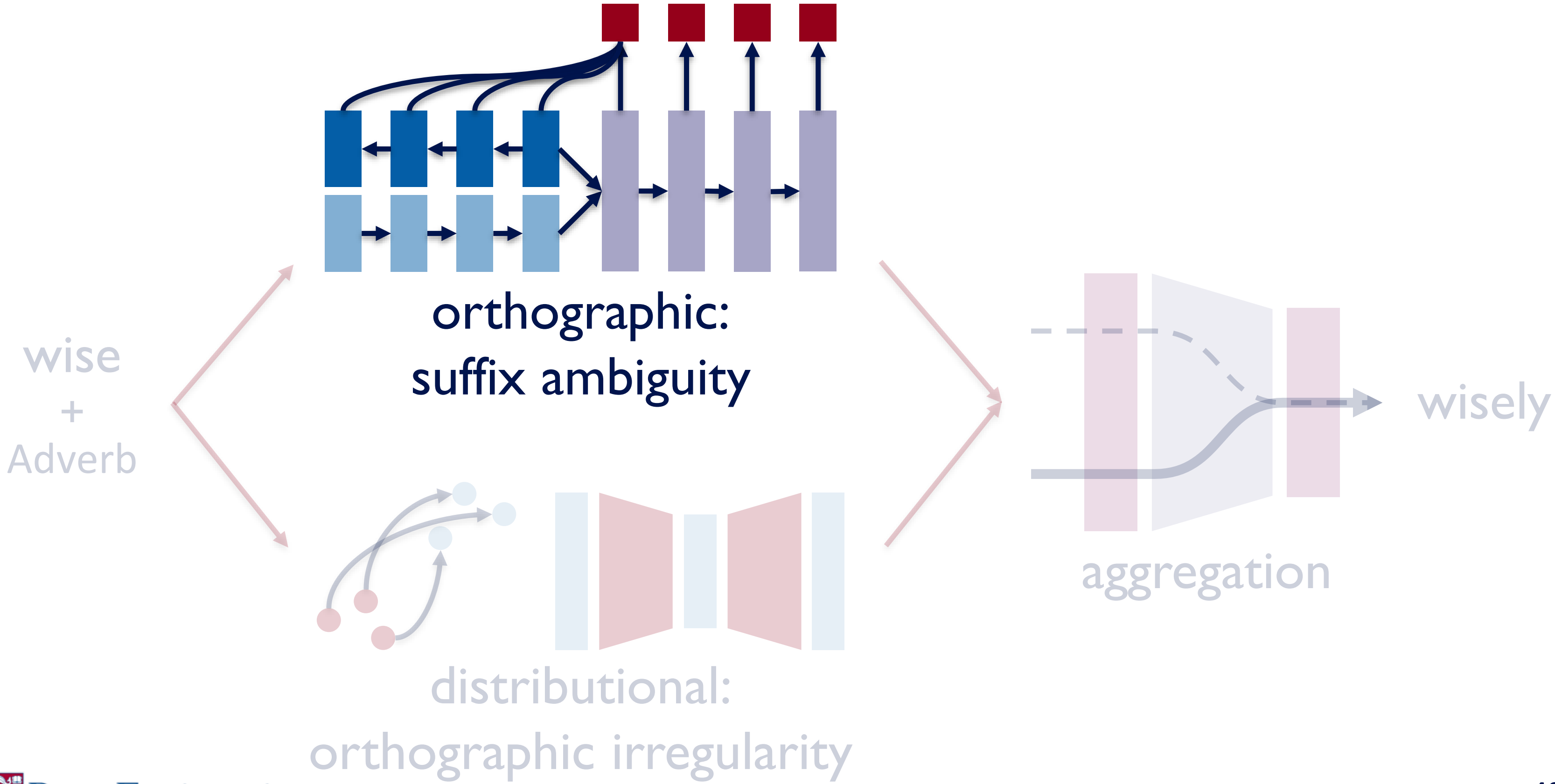
Reranking with Frequency Information

$\text{refute} \xrightarrow{\text{RESULT}}$

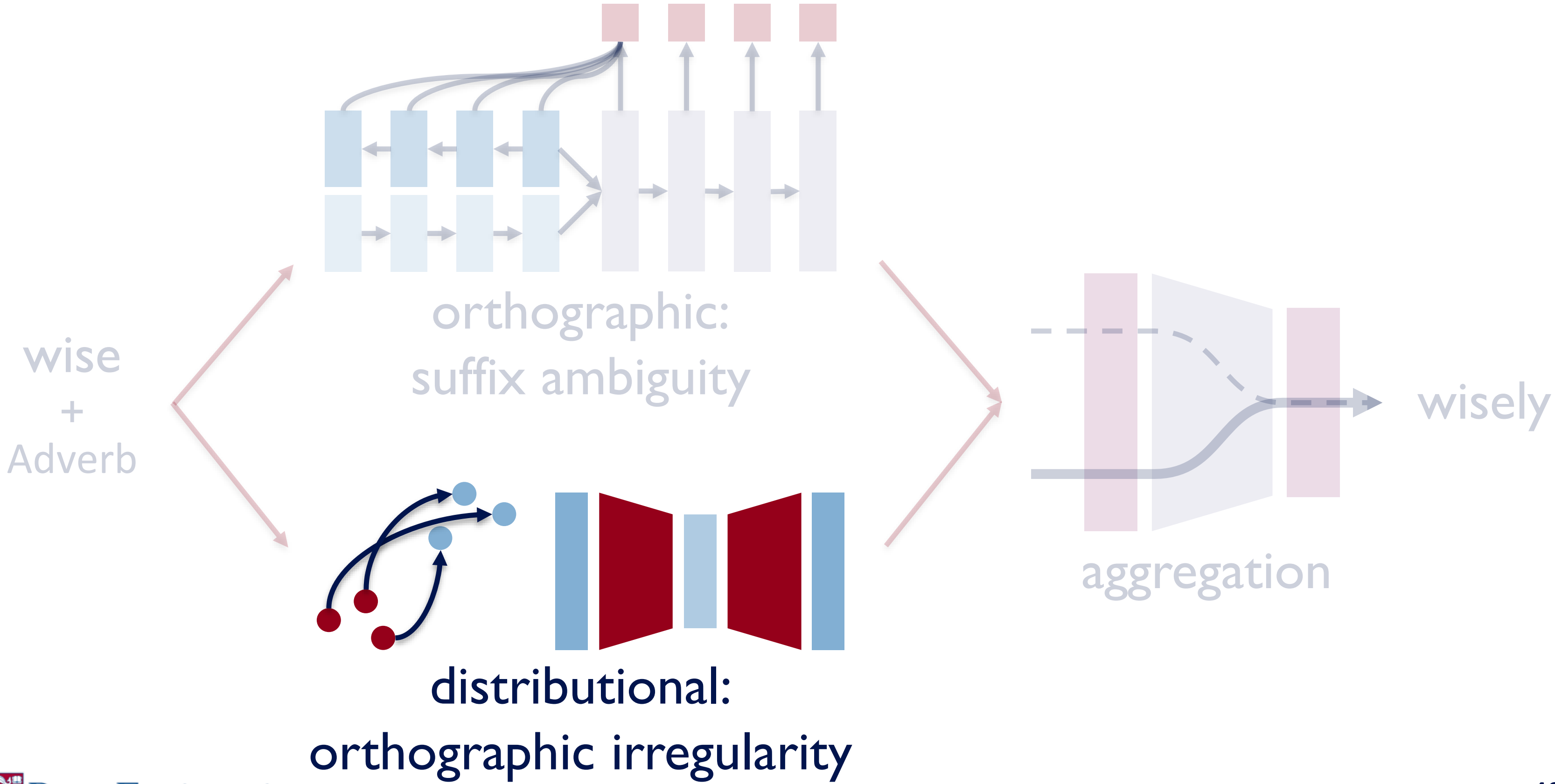
Model Output	Model Score	Log Corpus Freq	Reranker Output	Reranker Score
refution	-1.1	5.0	refutation	0.5
refutation	-1.2	14.3	refution	-0.9
refut	-4.8	7.4	refut	-0.9
refuty	-5.6	0.1	refuty	-0.9
refutat	-8.7	8.6	refutat	-0.9

Penn Engineering

Model Overview



Model Overview



Penn Engineering

Distributional Model

- Orthographic information can be unreliable
- Semantic transformation remains the same

Orthographic Irregularity

$\text{speak} \xrightarrow{\text{RESULT}} \text{spee}\textit{ch}$

$\text{creak} \xrightarrow{\text{RESULT}}$ *~~cree*ch*~~

creak*ing*

Penn Engineering

Distributional Model

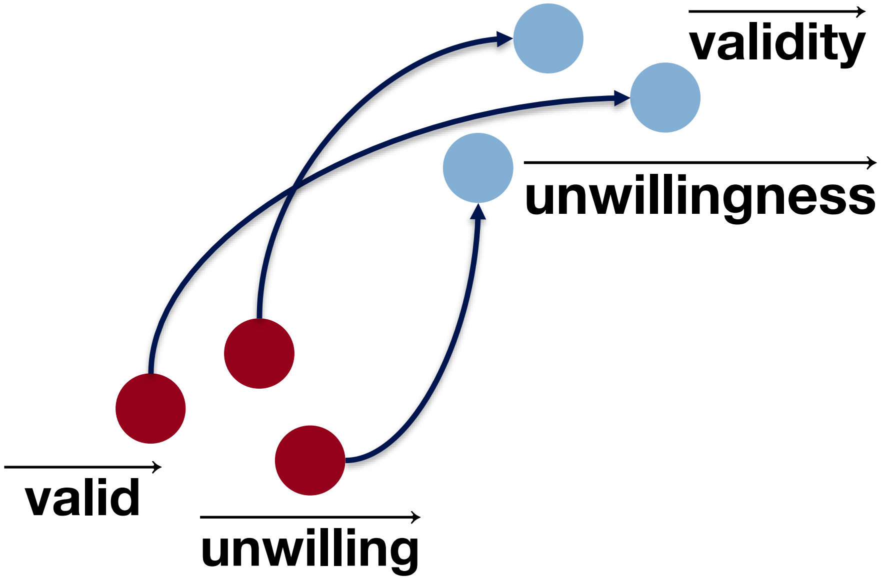
Intuition

$$\overrightarrow{\textbf{cats}} - \overrightarrow{\textbf{cat}} + \overrightarrow{\textbf{ox}} \sim \overrightarrow{\textbf{oxen}}$$

Distributional Model

Intuition

$$\overrightarrow{\textbf{cats}} - \overrightarrow{\textbf{cat}} + \overrightarrow{\textbf{ox}} \sim \overrightarrow{\textbf{oxen}}$$

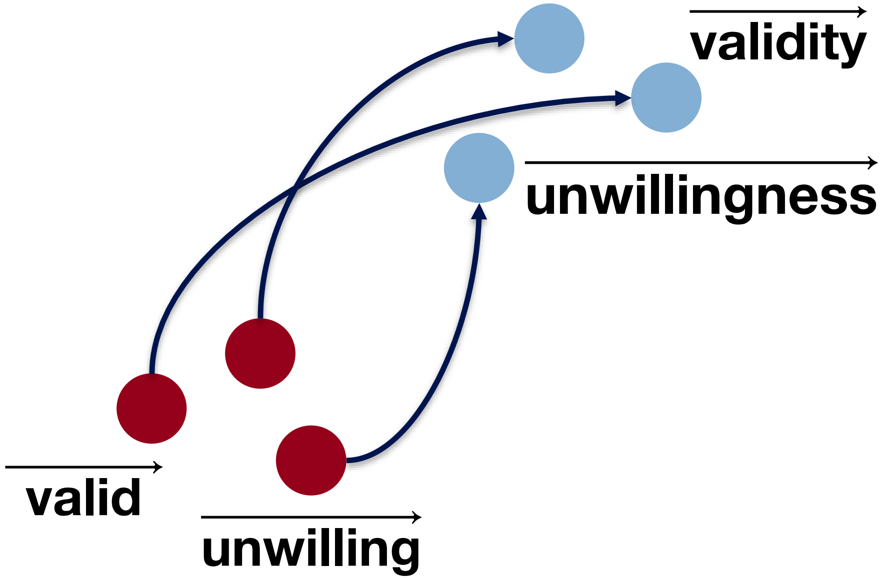


Distributional Model

Intuition

$$\overrightarrow{\textbf{cats}} - \overrightarrow{\textbf{cat}} + \overrightarrow{\textbf{ox}} \sim \overrightarrow{\textbf{oxen}}$$

Learn non-linear function per transformation



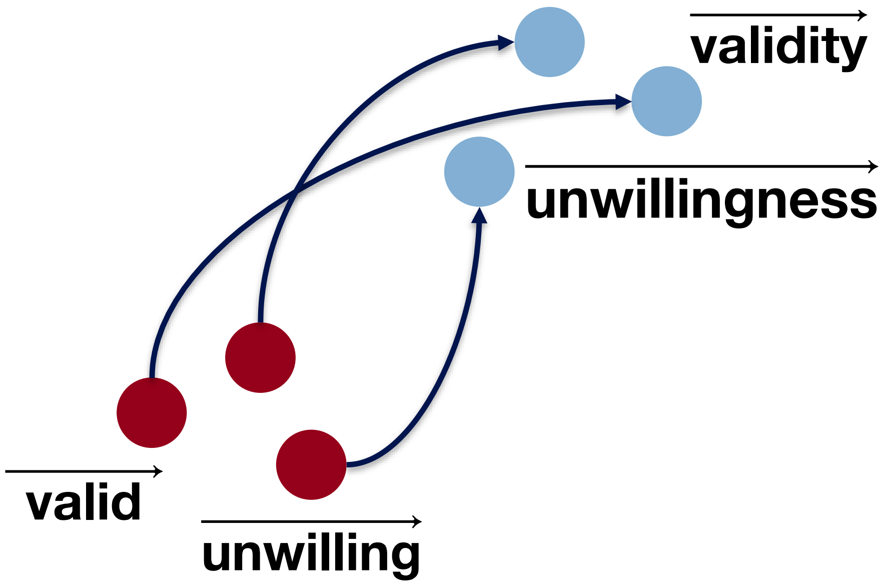
Distributional Model

Intuition

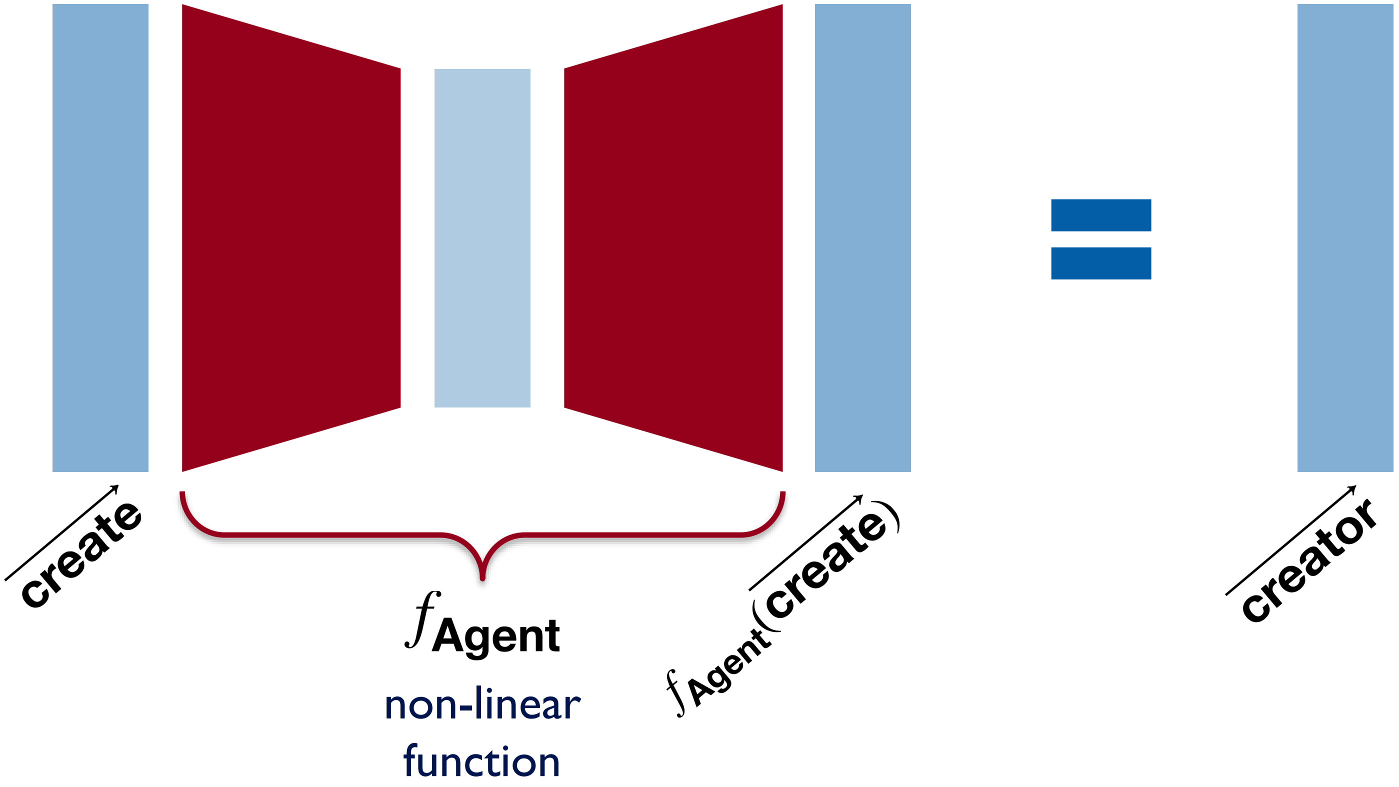
$$\overrightarrow{\textbf{cats}} - \overrightarrow{\textbf{cat}} + \overrightarrow{\textbf{ox}} \sim \overrightarrow{\textbf{oxen}}$$

Learn non-linear function per transformation

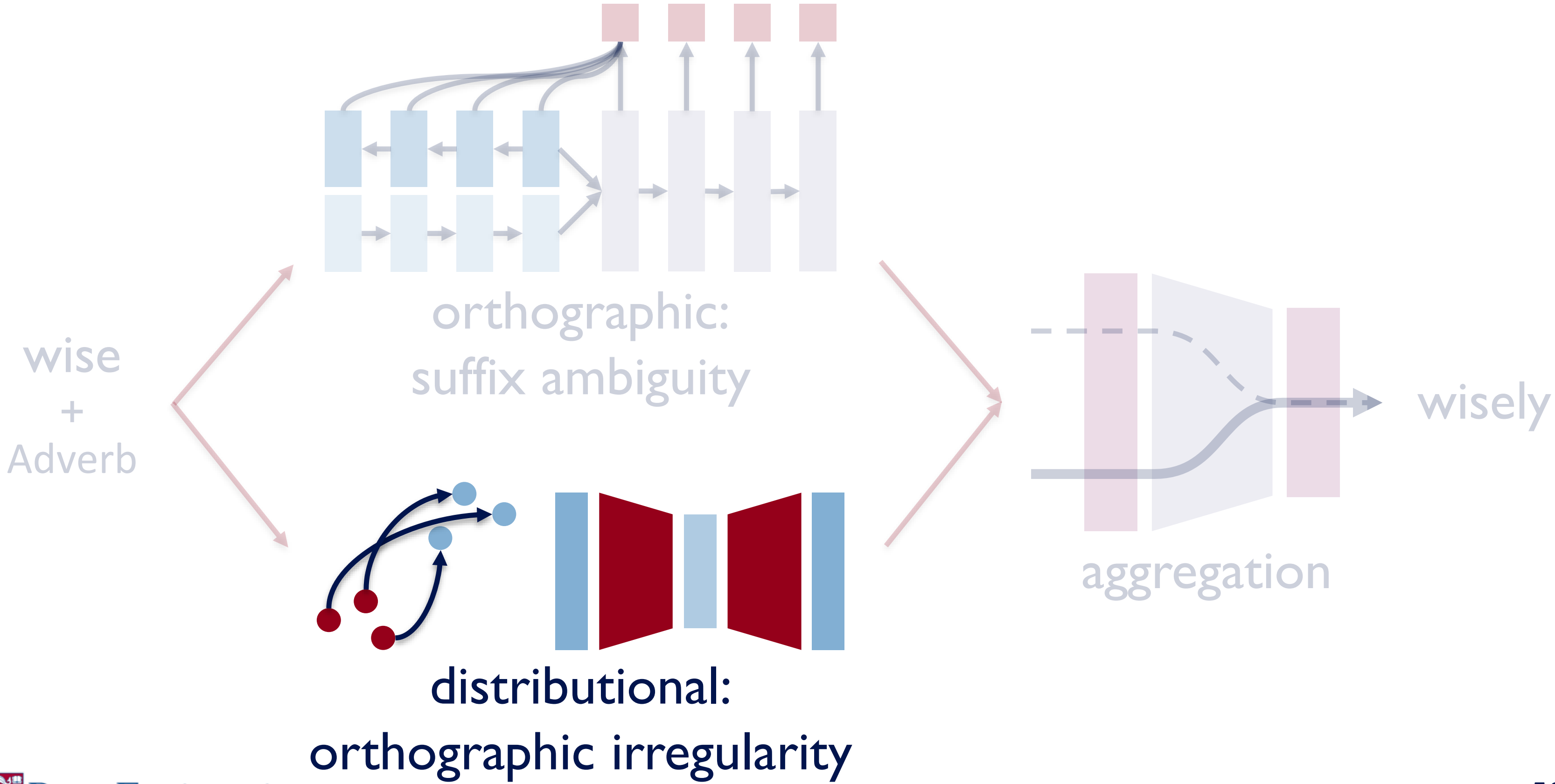
Independent of orthography



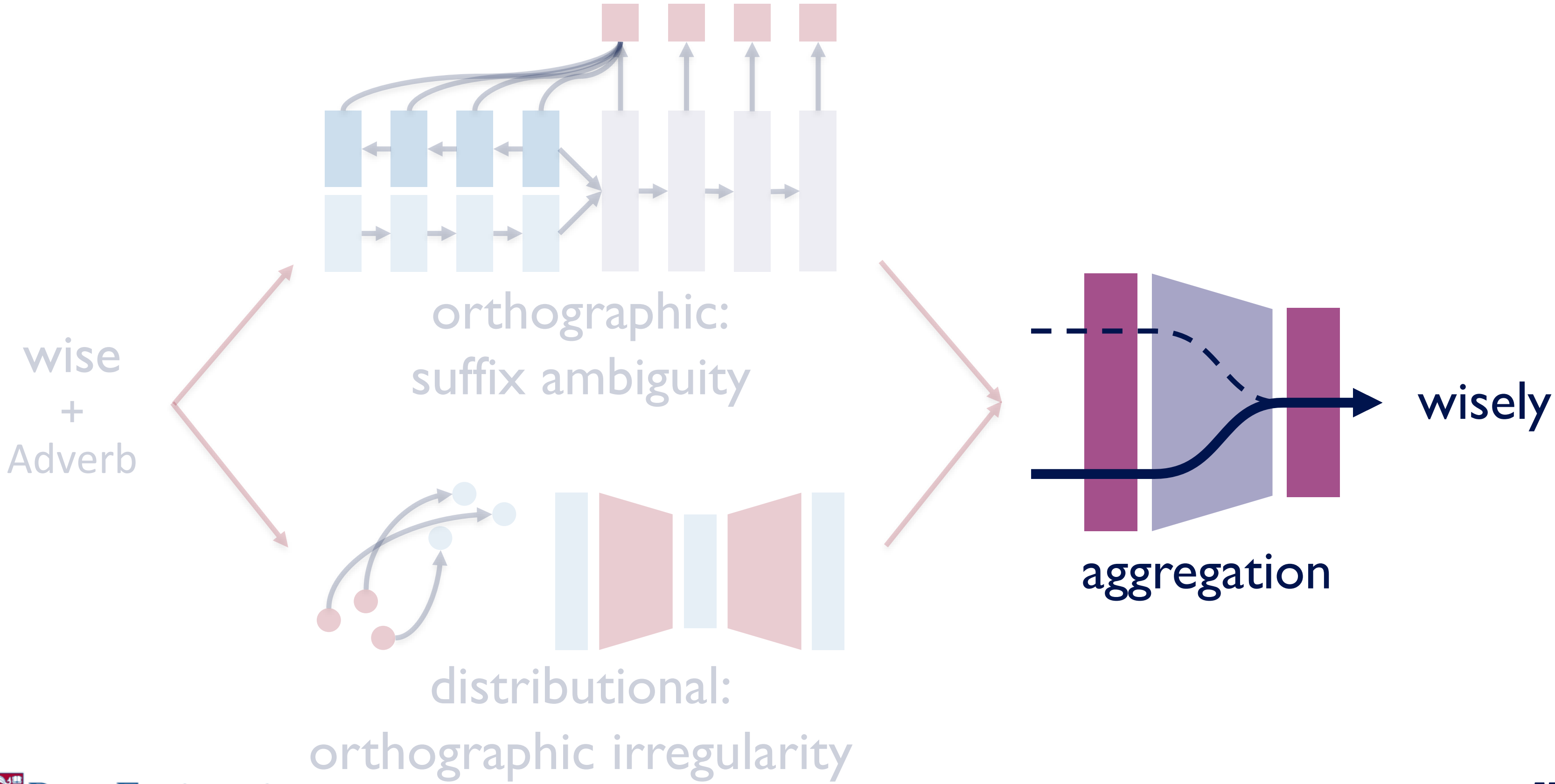
Distributional Model



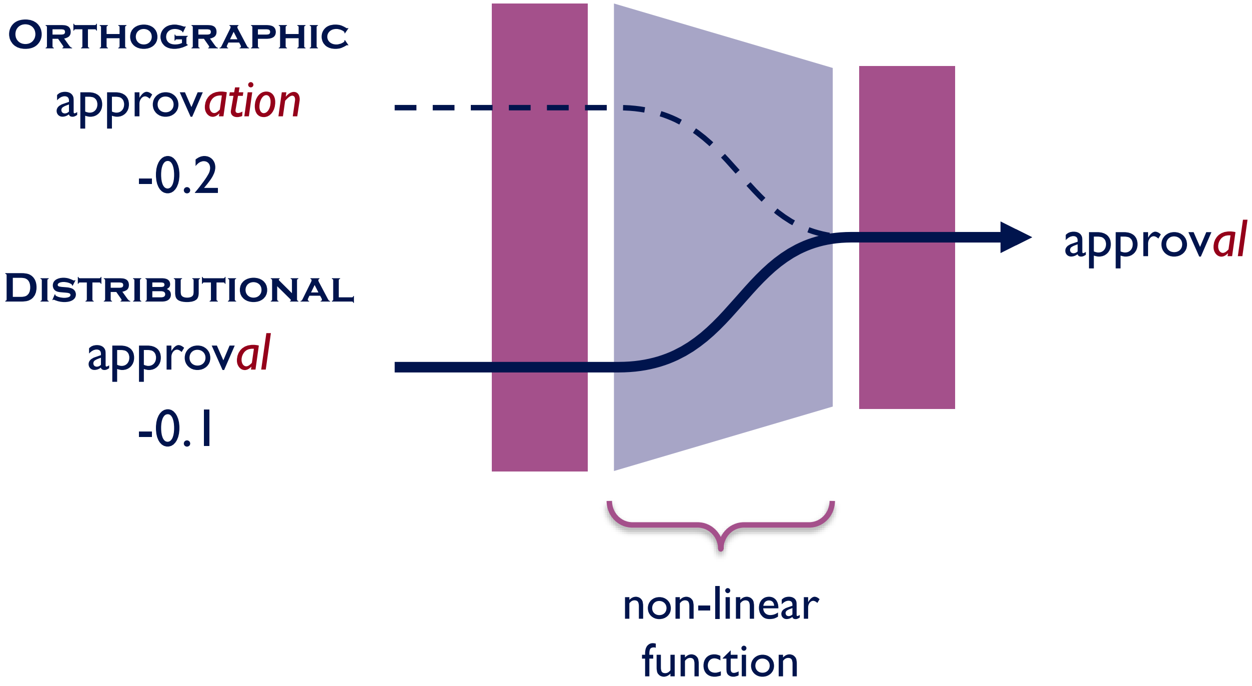
Model Overview



Model Overview



Aggregation Model



Aggregation Model

Ortho	Score	Distributional	Score
approvation	-0.9	approval	-0.6
bankruption	-0.3	bankruptcy	-0.8
expertly	-0.5	expertly	-1.1
stroller	-0.8	strolls	-0.9

Aggregation Model

Ortho	Score	Distributional	Score
approvation	-0.9	approval	-0.6
bankruption	-0.3	bankruptcy	-0.8
expertly	-0.5	expertly	-1.1
stroller	-0.8	strolls	-0.9

Aggregation Model

Ortho	Score	Distributional	Score	Aggregation Selection
approvation	-0.9	approval	-0.6	approval
bankruption	-0.3	bankruptcy	-0.8	bankruption
expertly	-0.5	expertly	-1.1	expertly
stroller	-0.8	strolls	-0.9	stroller

Experiments

Dataset

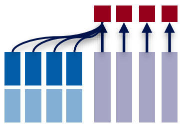
Cotterell et al. 2017

Transformation	Count	Example
Adverb	1715	wise → wisely
Result	1251	simulate → simulation recite → recital overstate → overstatement
Agent	801	yodel → yodeler survive → survivor
Nominal	354	intense → intensity effective → effectiveness pessimistic → pessimism

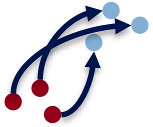
Experiment Details

- 30 random restarts
- Token information: Google Book NGrams
  - 360k unigram types
  - Token counts aggregated
- Google News pre-trained word embeddings
- Evaluation: full-token match accuracy

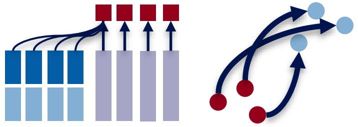
Results Legend



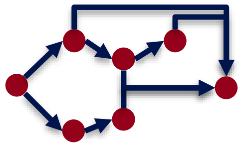
Seq2Seq



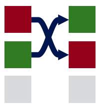
Distributional



Aggregation

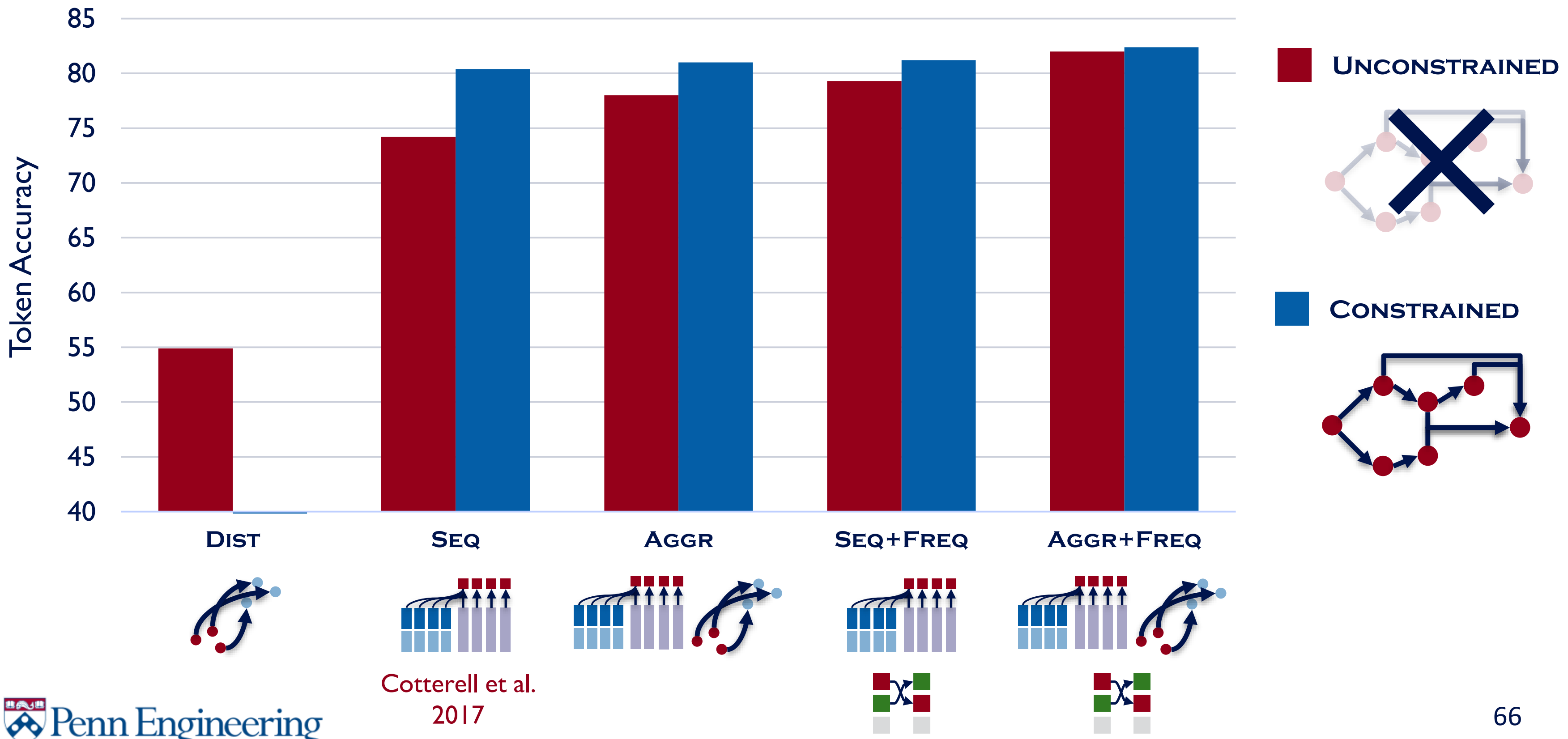


Dictionary-Constrained Decoding



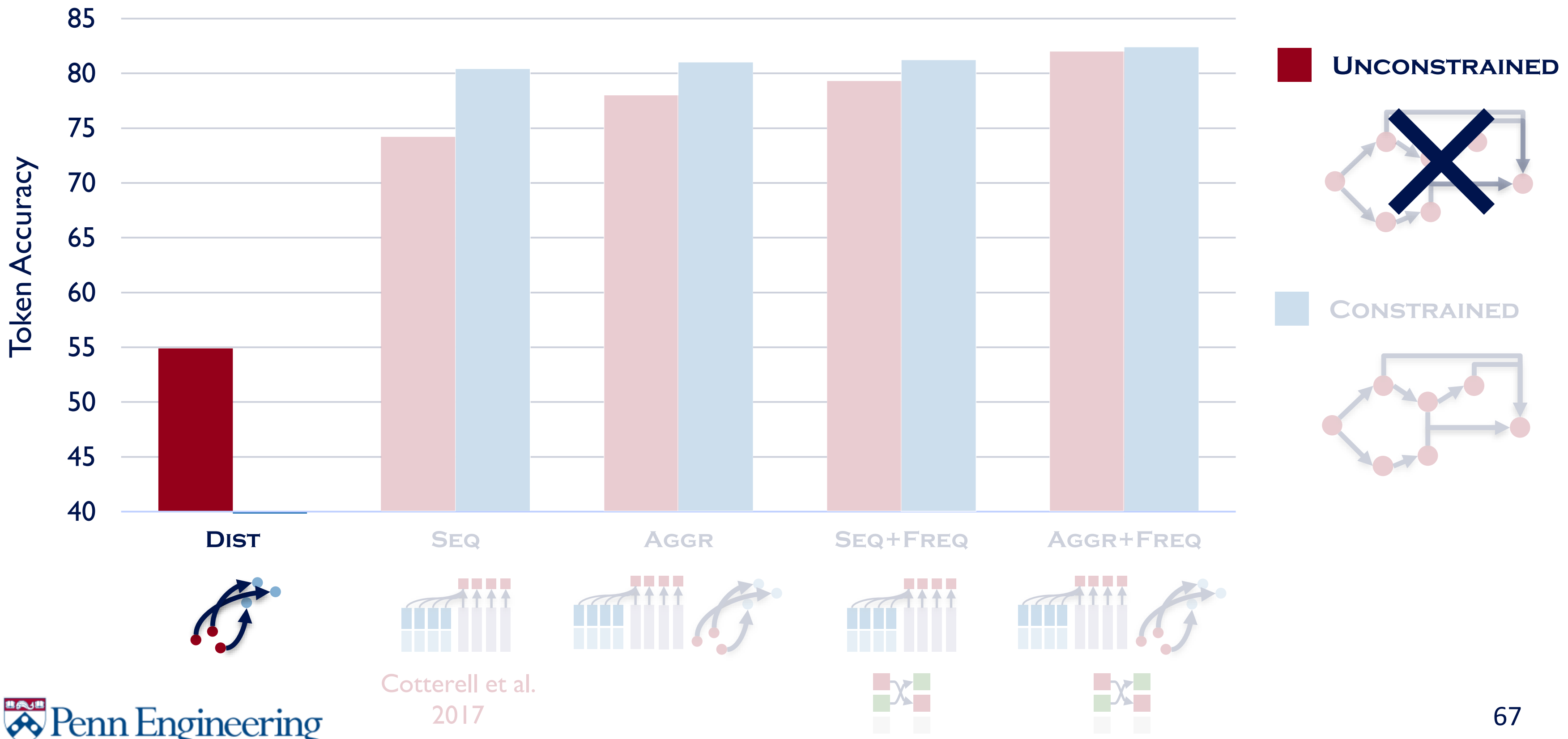
Frequency-Based Reranking

Results



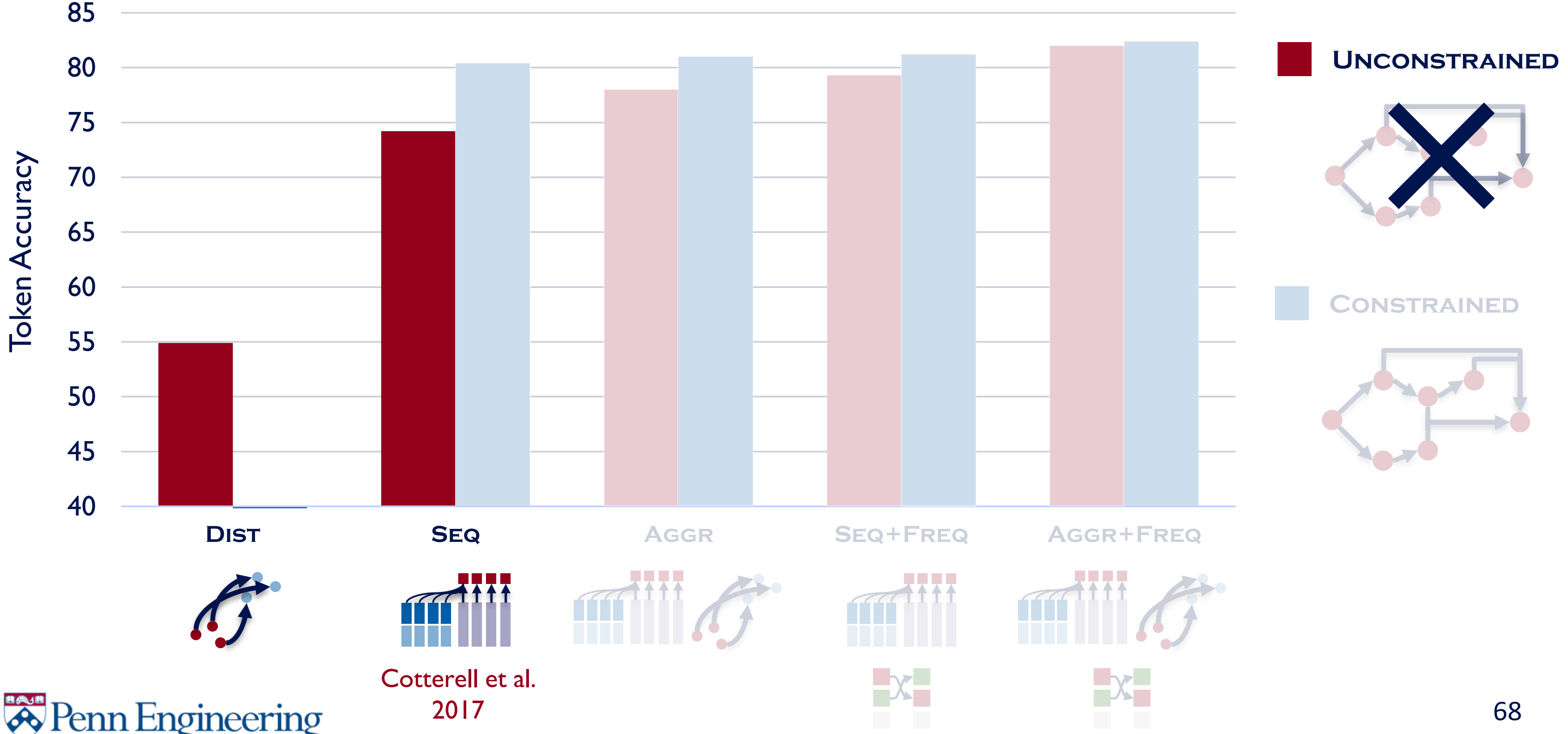
Penn Engineering

Results



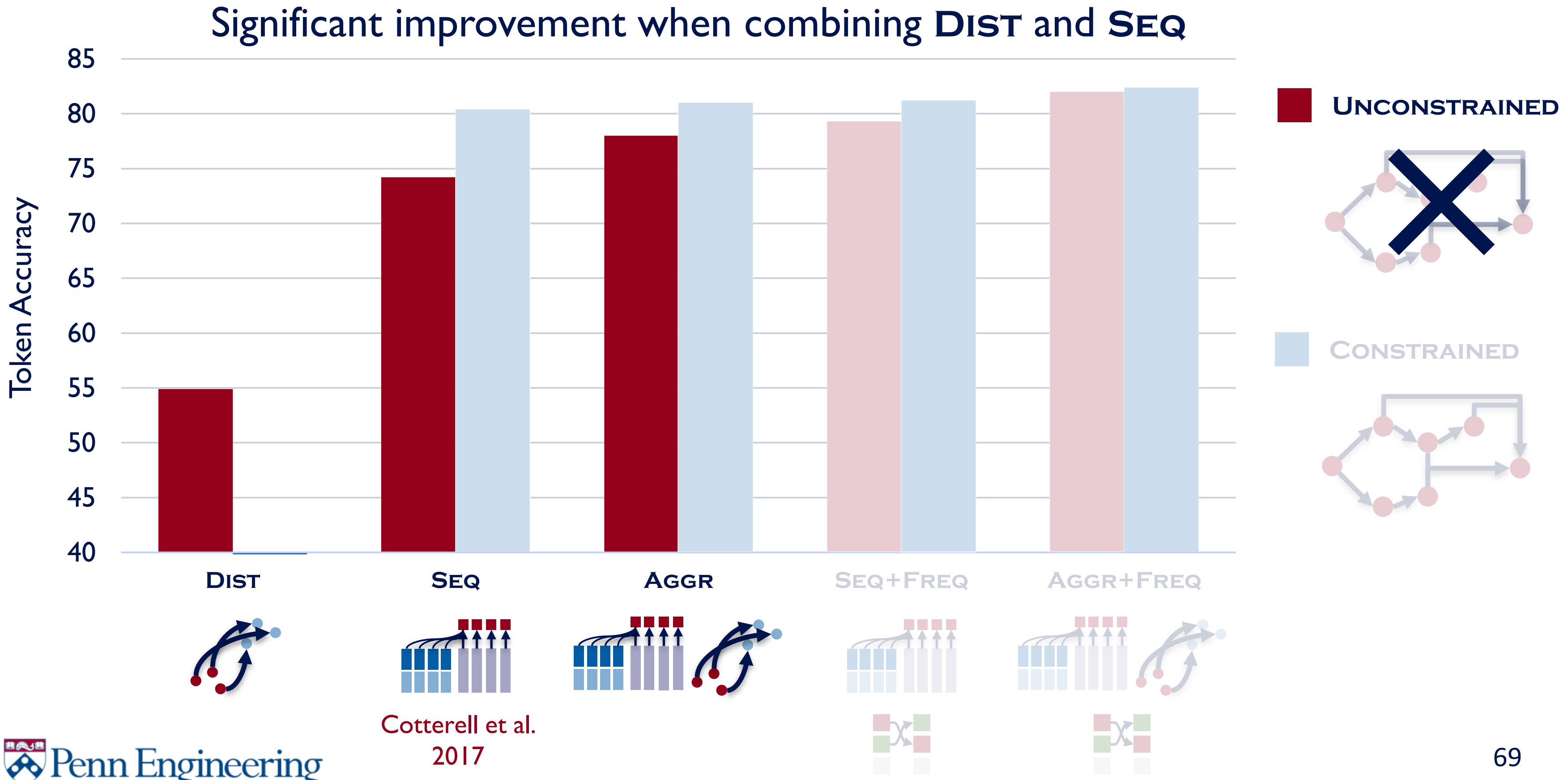
85
80
75
70
65
60
55
50
45
40
Token Accuracy
Dist
Seq
Aggr
Seq+Freq
Aggr+Freq
Unconstrained
Constrained
Cotterell et al. 2017

Results



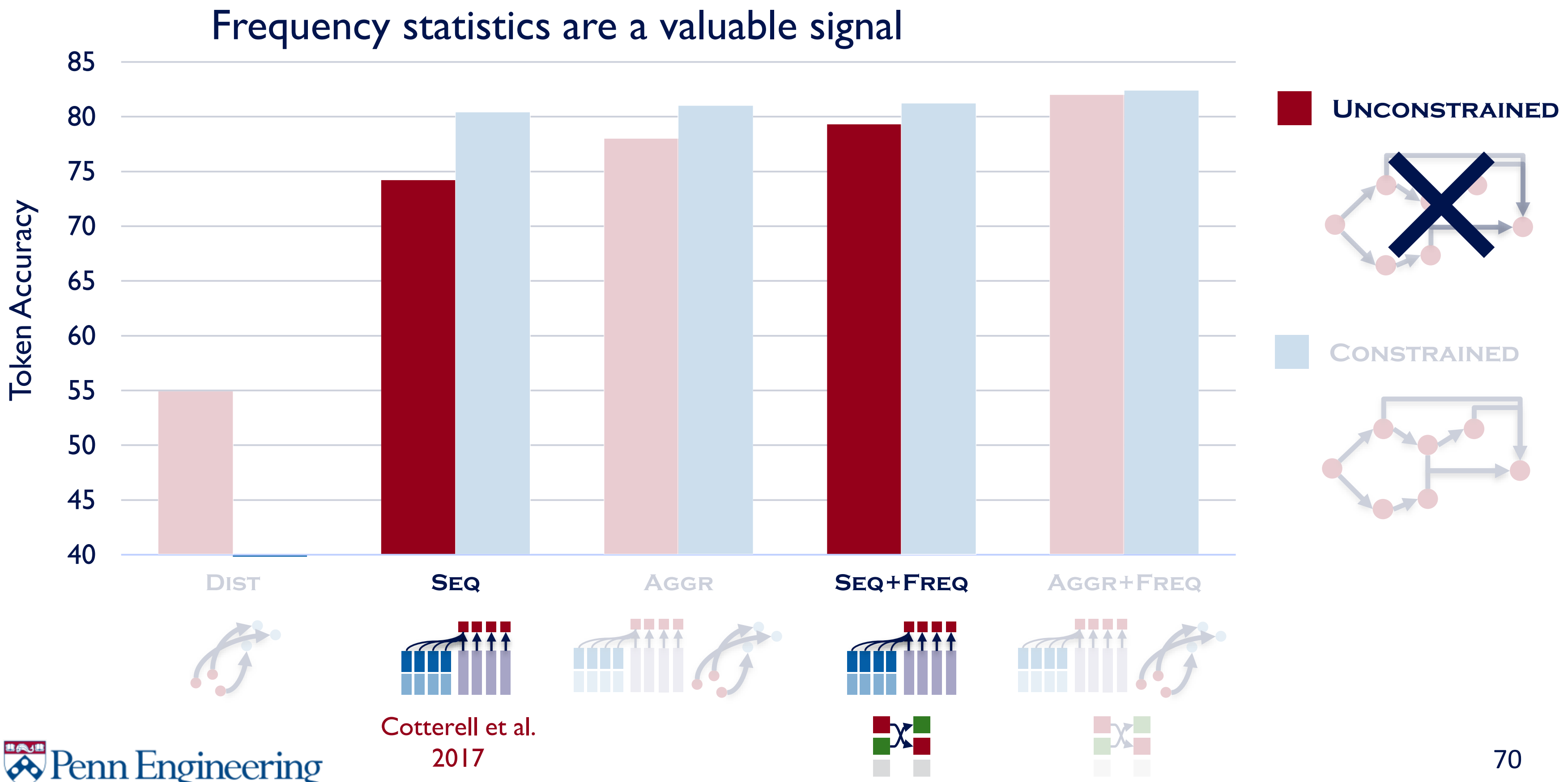
Cotterell et al. 2017

Results



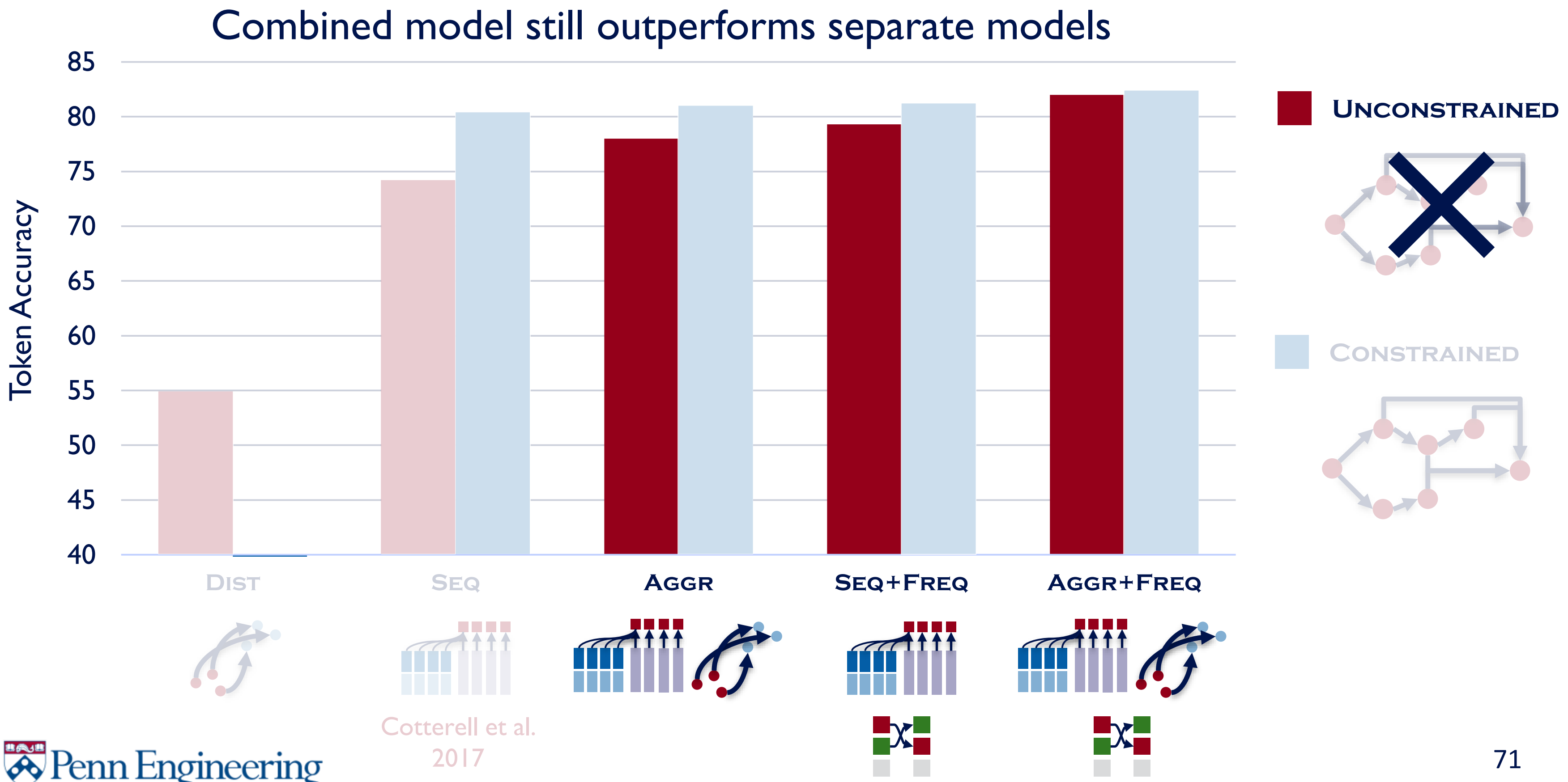
Significant improvement when combining DIST and SEQ
Token Accuracy
85
80
75
70
65
60
55
50
45
40
DIST
SEQ
AGGR
SEQ+FREQ
AGGR+FREQ
UNCONSTRAINED
CONSTRAINED
Cotterell et al.
2017

Results



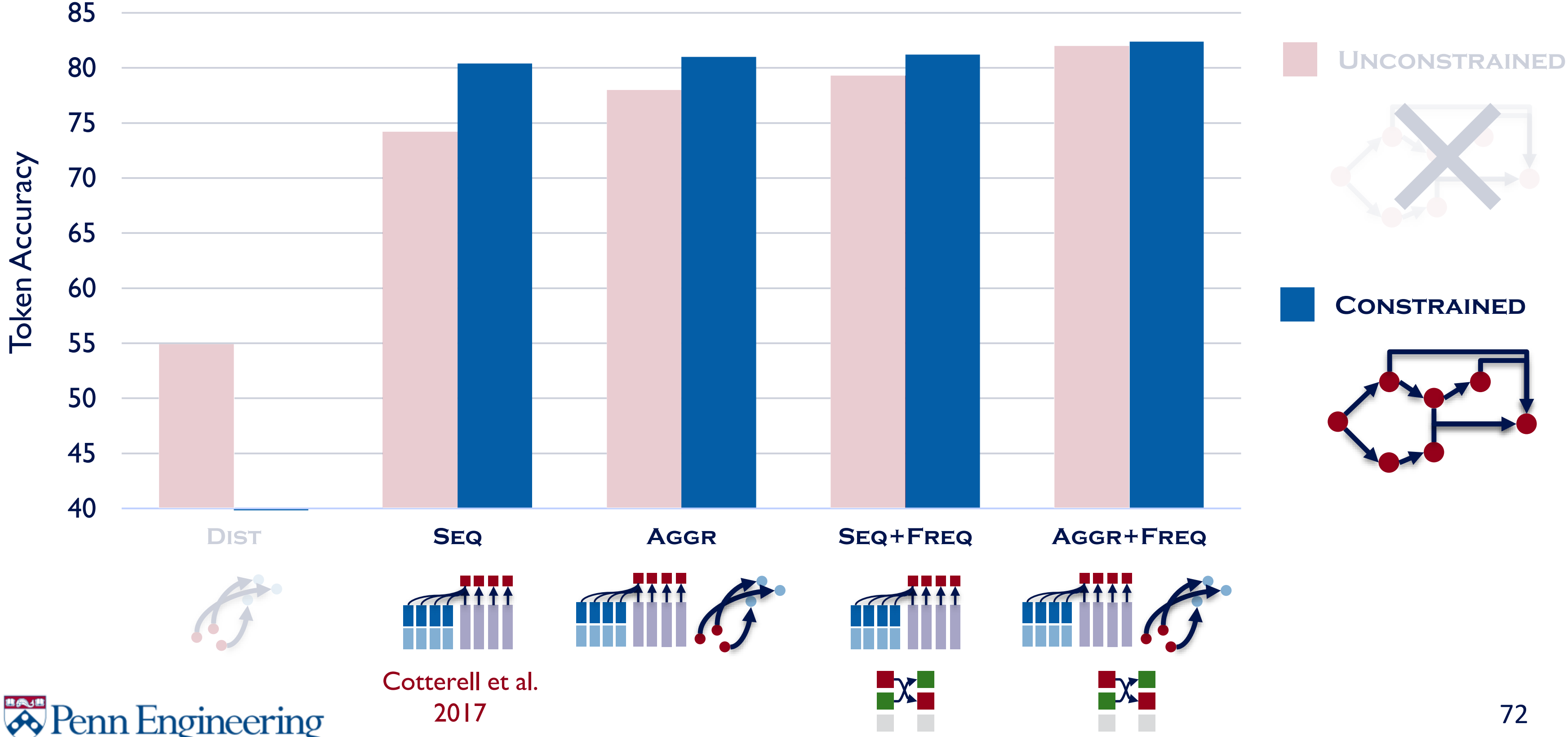
Frequency statistics are a valuable signal
Token Accuracy
85
80
75
70
65
60
55
50
45
40
DIST
SEQ
AGGR
SEQ+FREQ
AGGR+FREQ
Cotterell et al. 2017
UNCONSTRAINED
CONSTRAINED

Results



Combined model still outperforms separate models
Token Accuracy
85
80
75
70
65
60
55
50
45
40
DIST
SEQ
AGGR
SEQ+FREQ
AGGR+FREQ
Cotterell et al. 2017
UNCONSTRAINED
CONSTRAINED

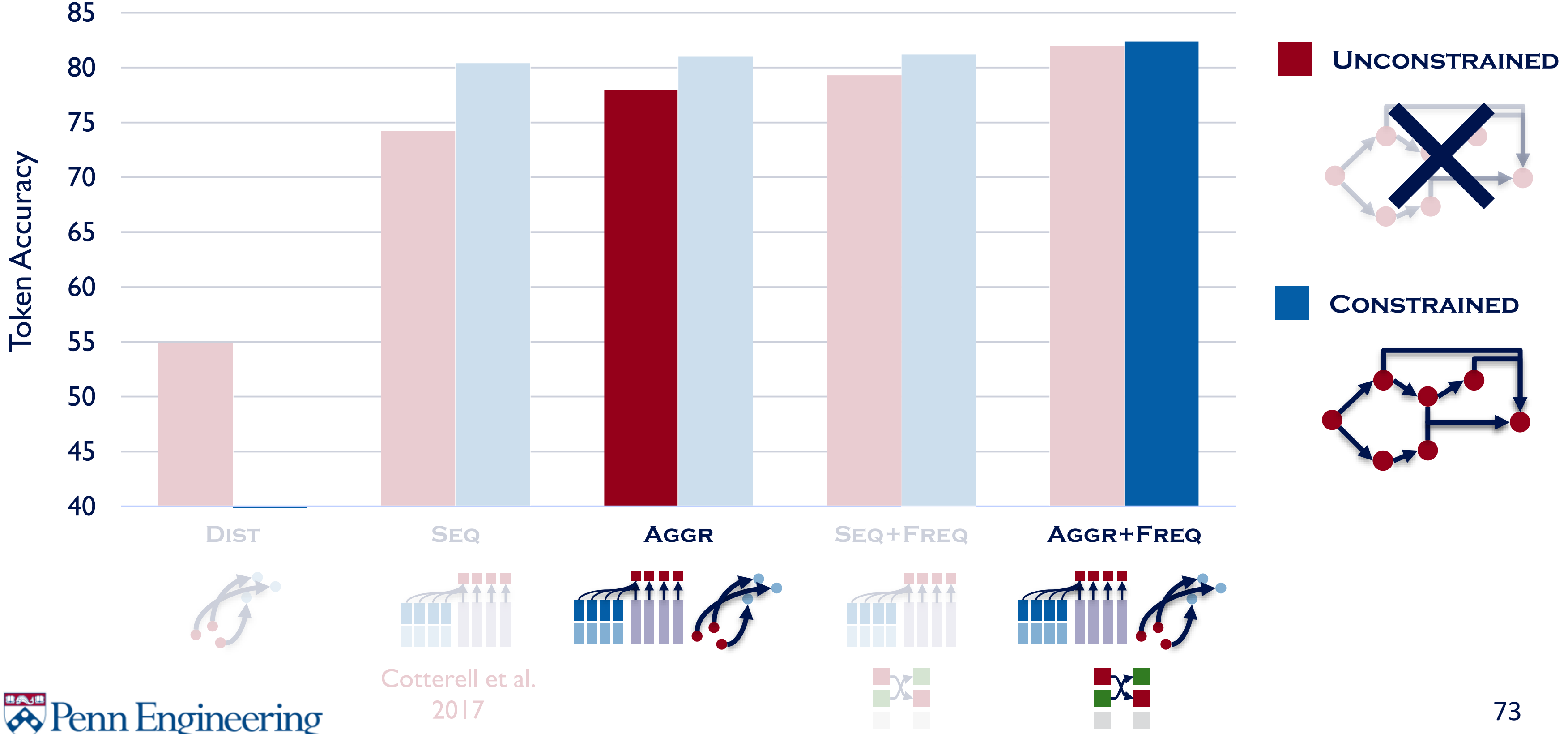
Results



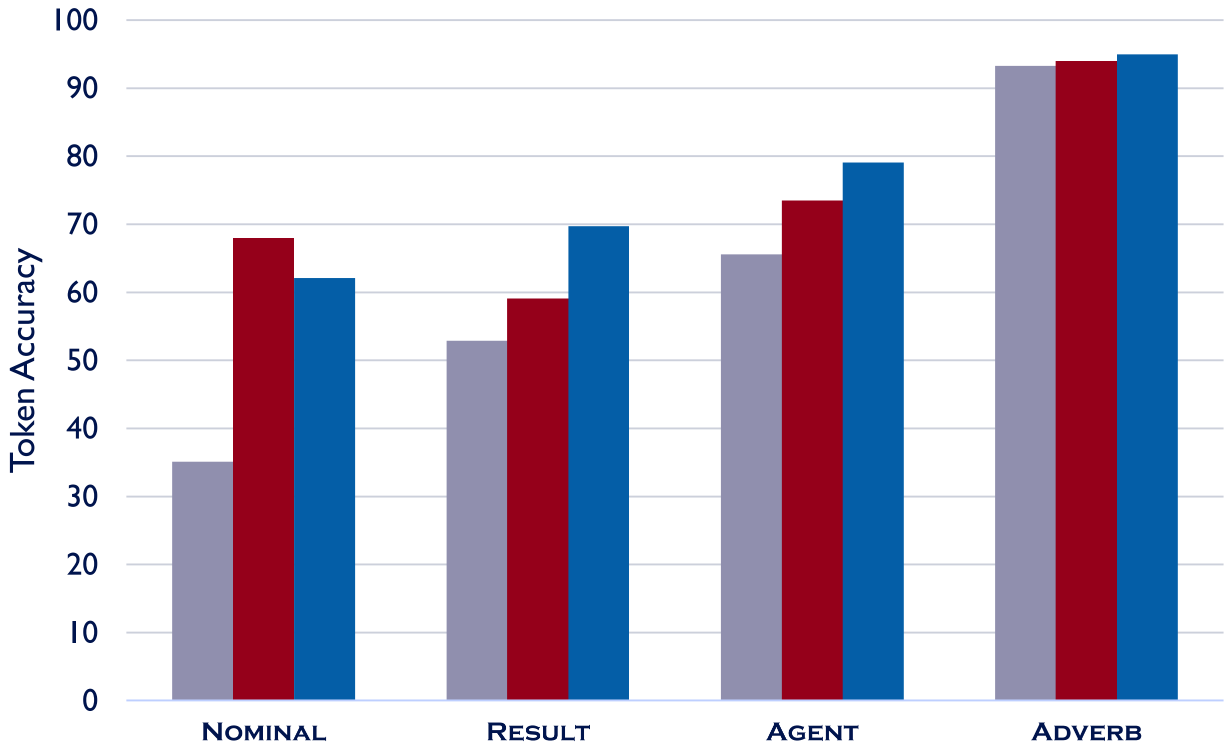
Cotterell et al.
2017

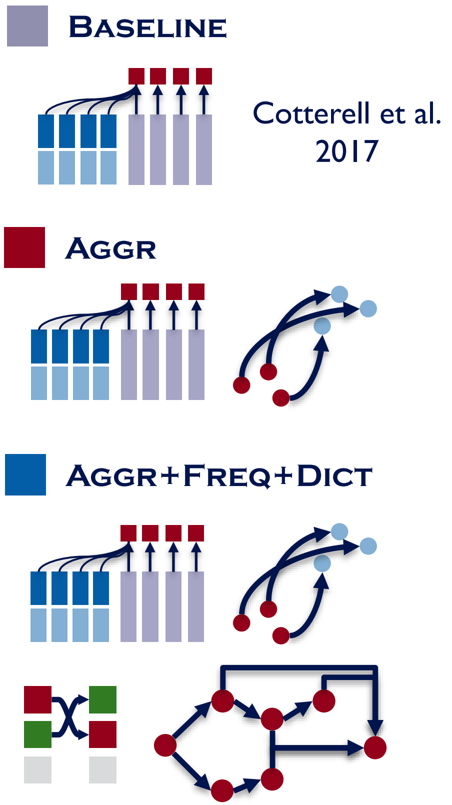
Results

22% and 37% relative error reductions over SEQ

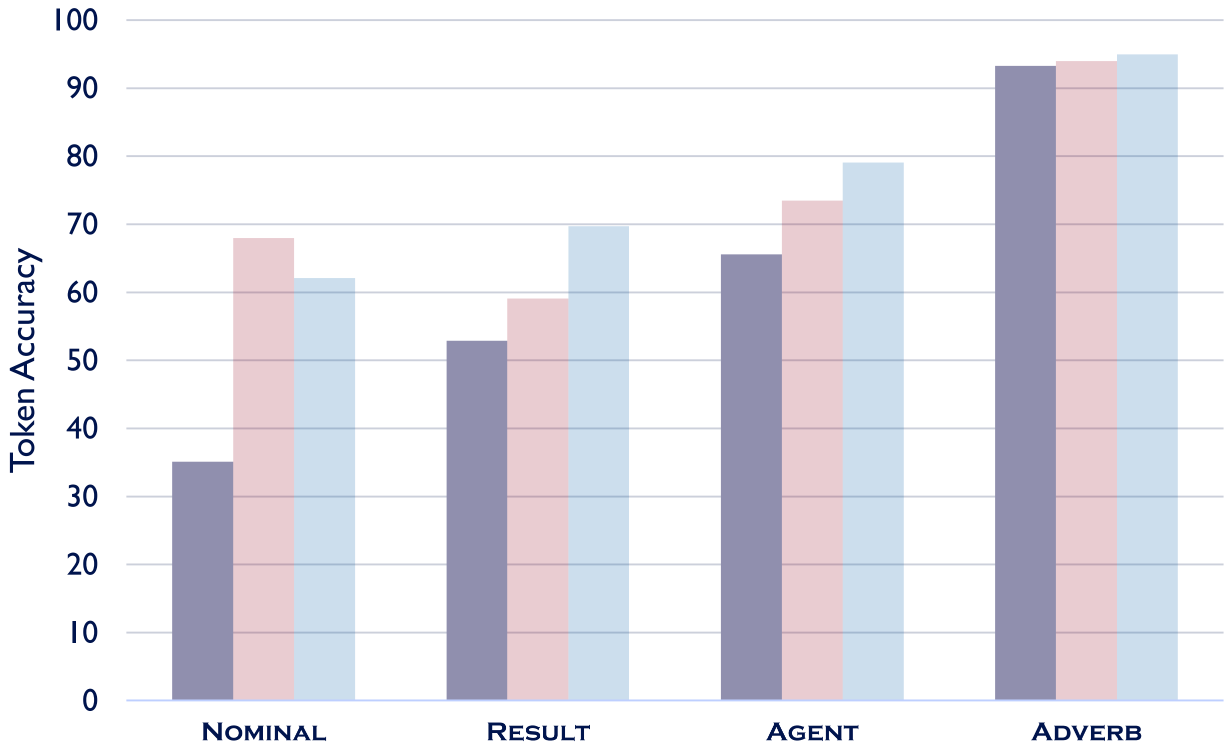


Results by Transformation

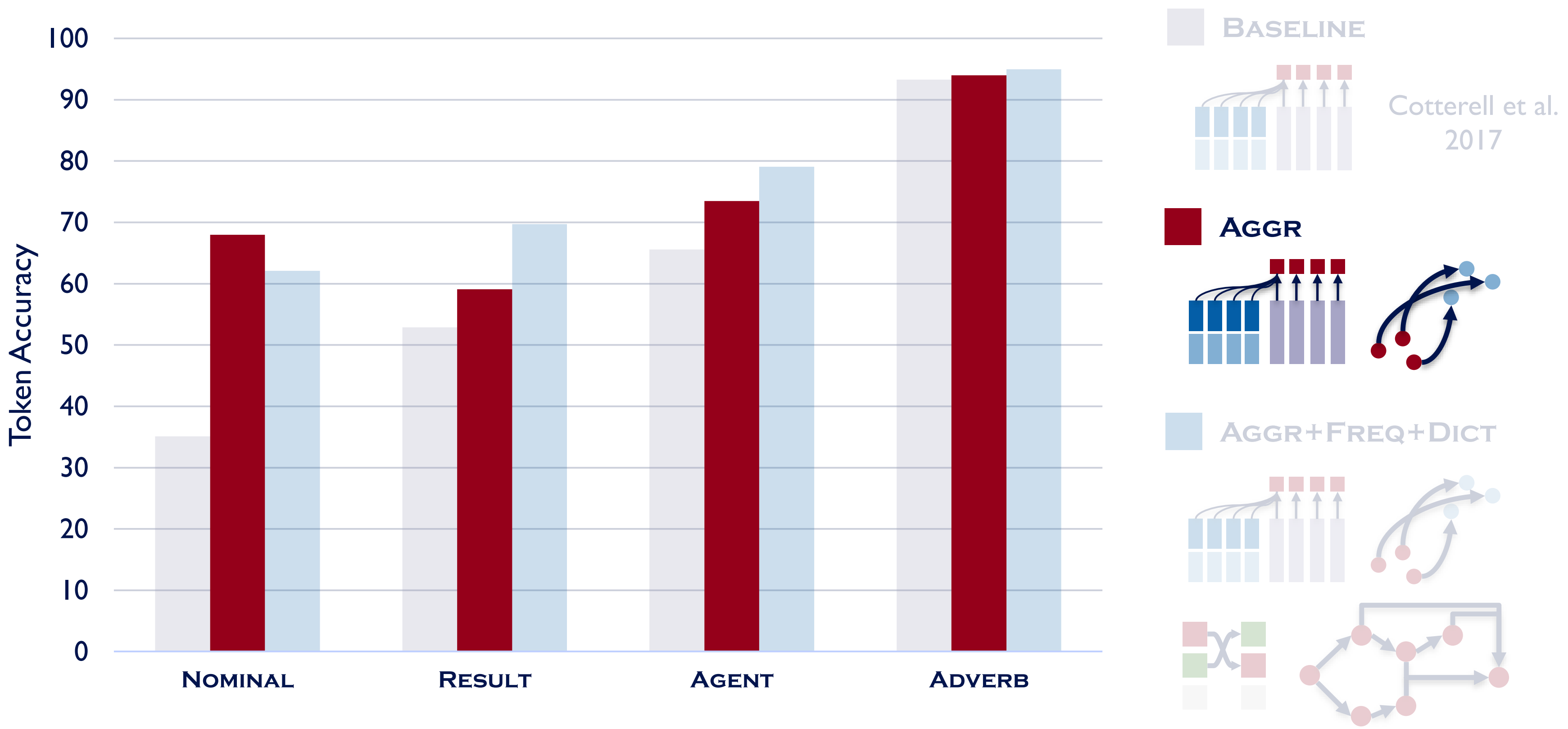




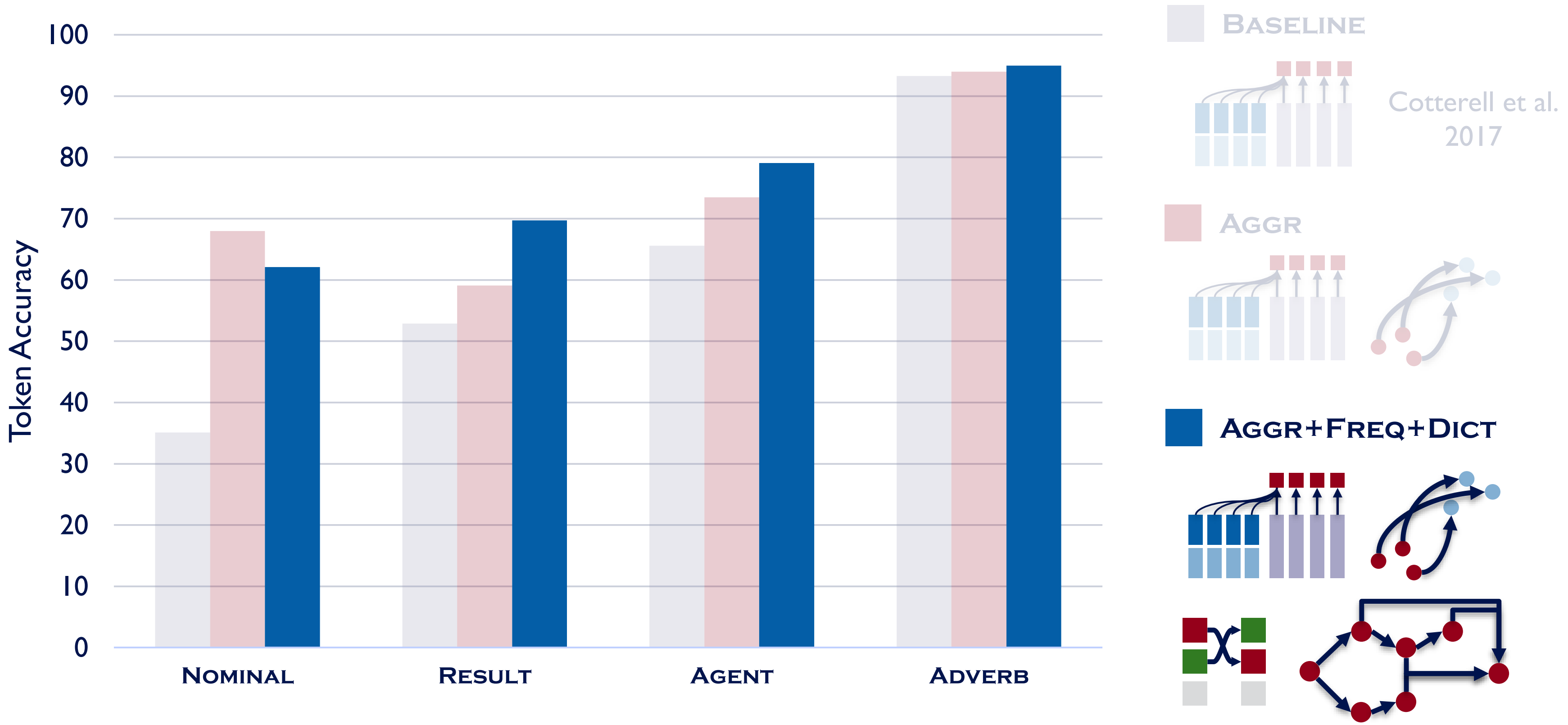
Results by Transformation



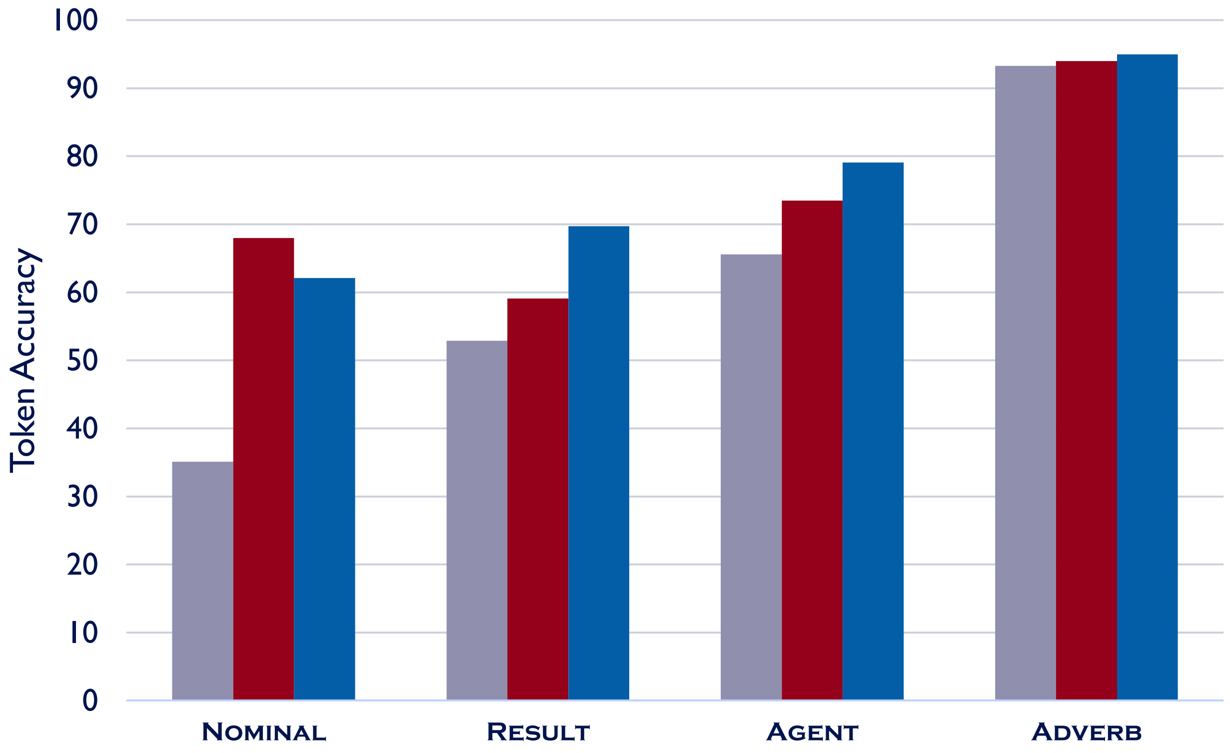
Results by Transformation

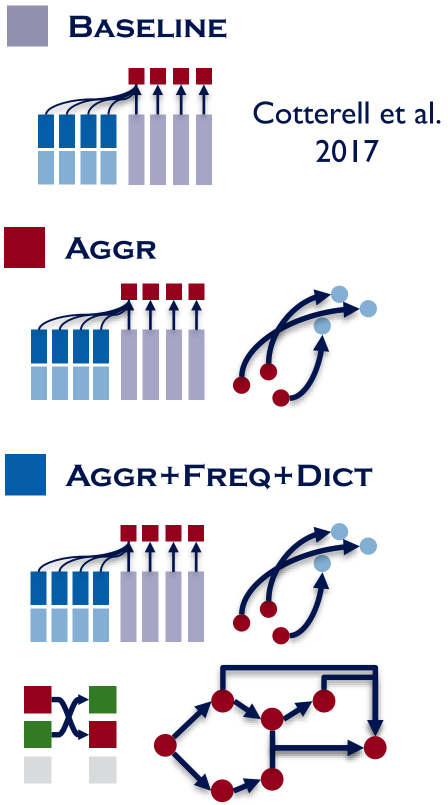


Results by Transformation

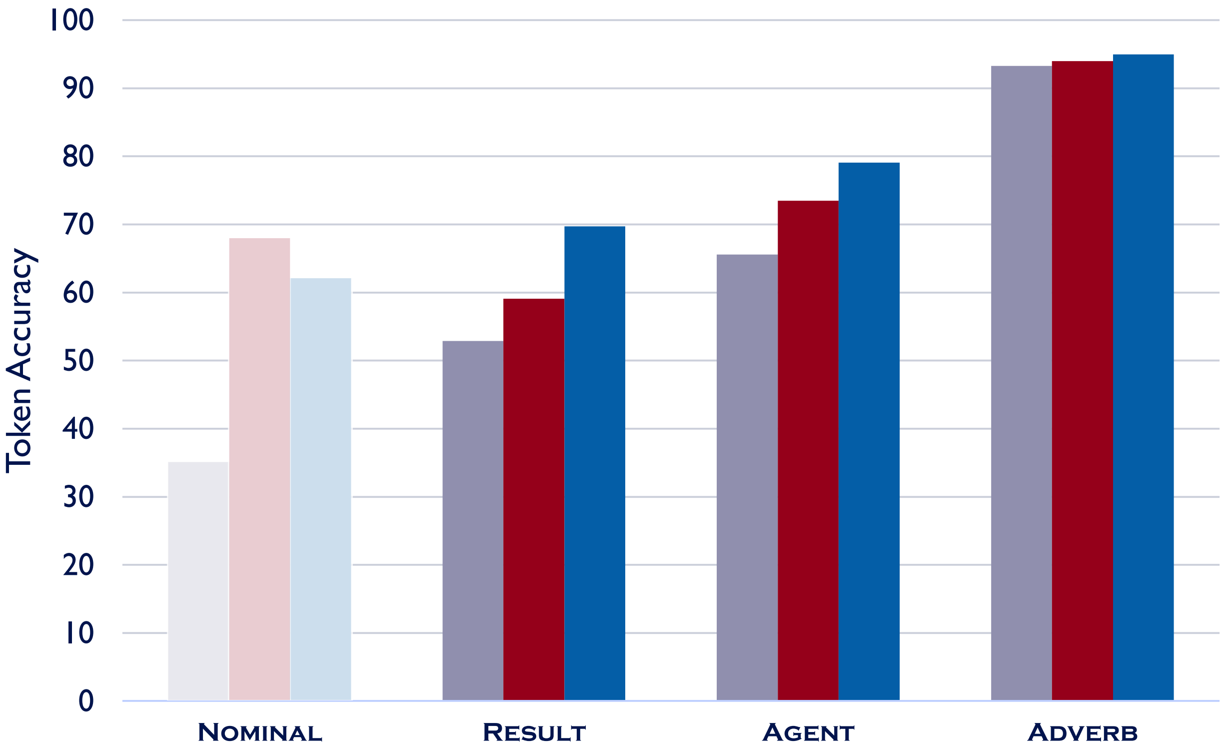


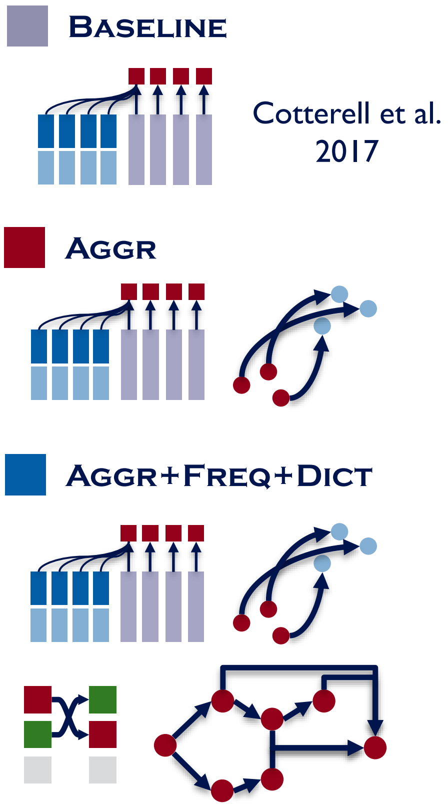
Results by Transformation



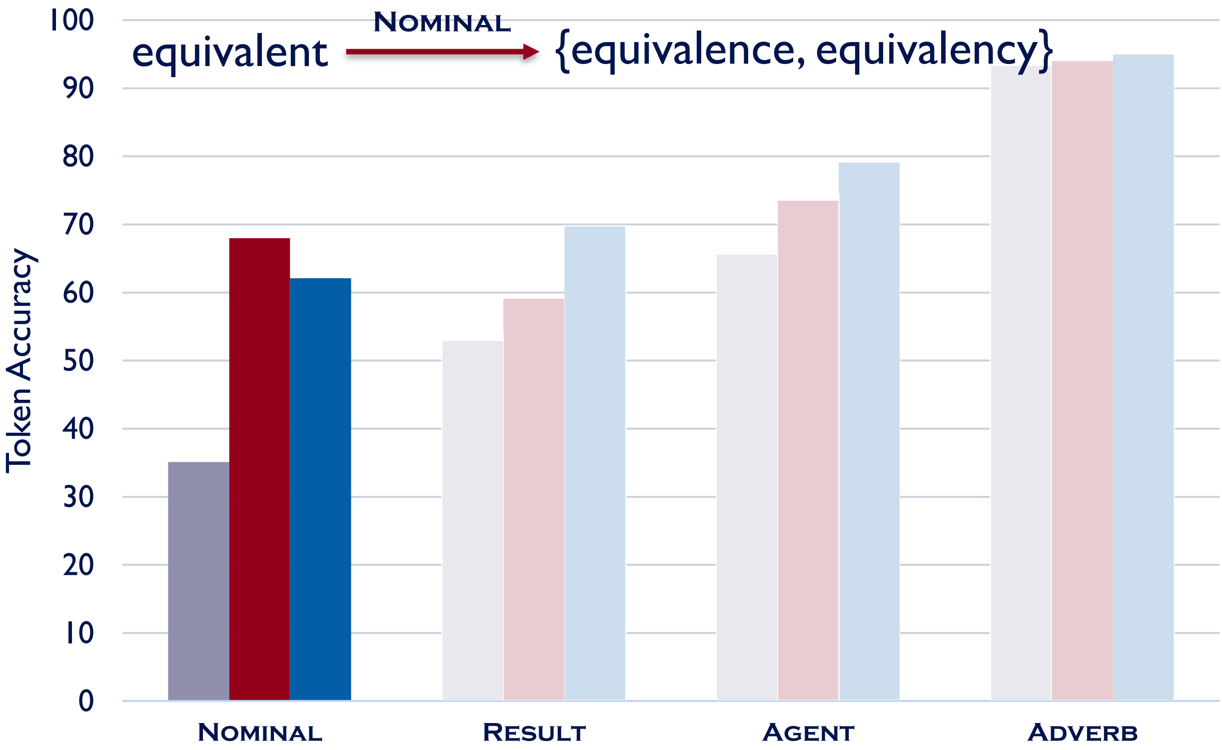


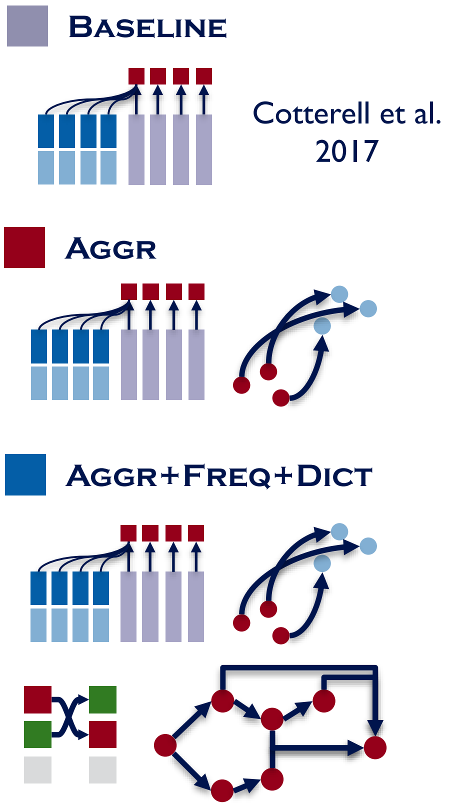
Results by Transformation



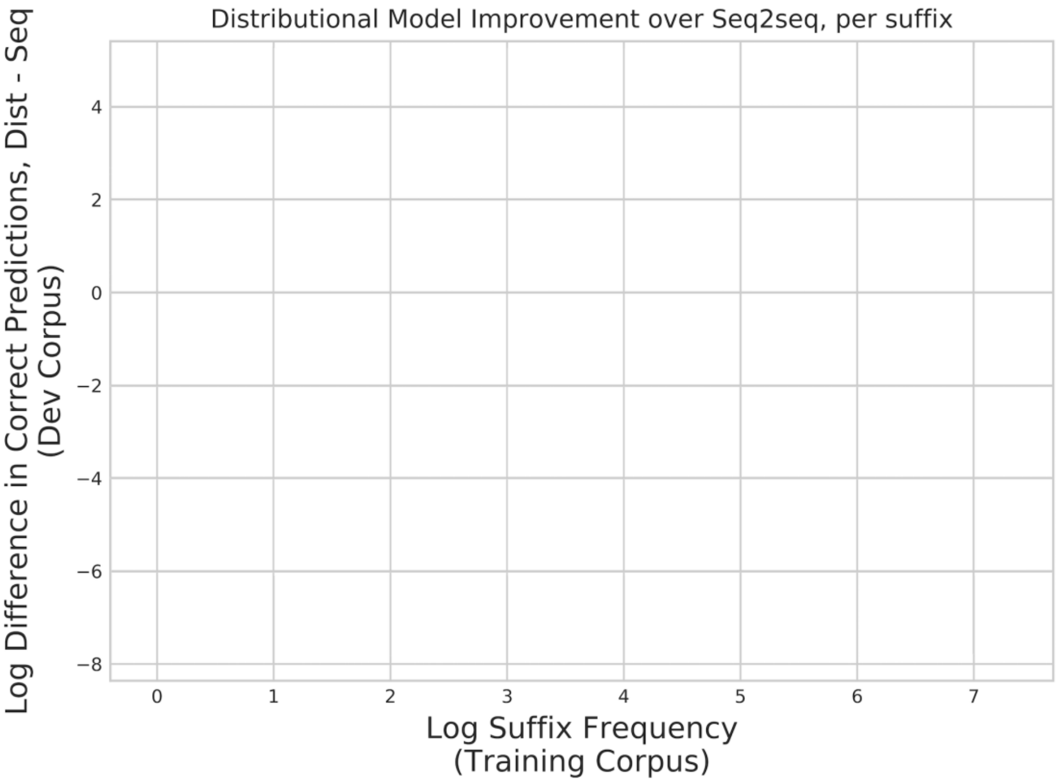


Results by Transformation



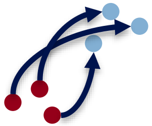


What does each model do well?

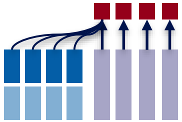


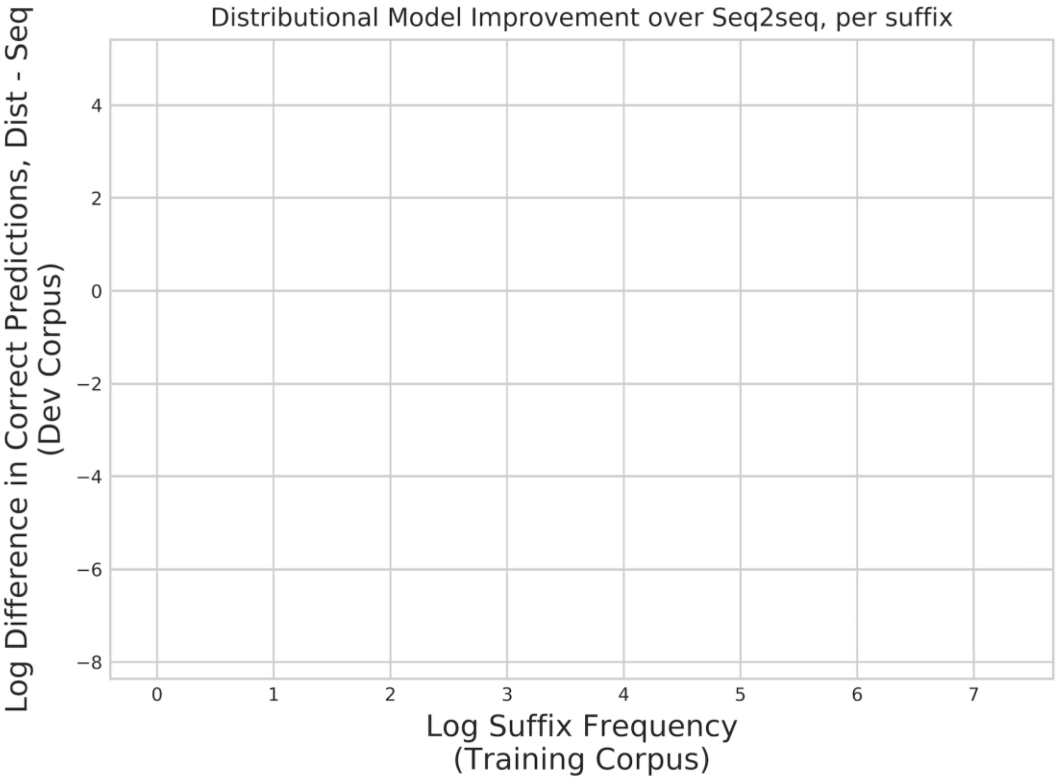
What does each model do well?

Distributional does best



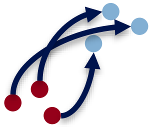
Orthographic does best



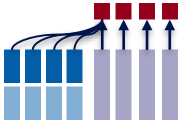


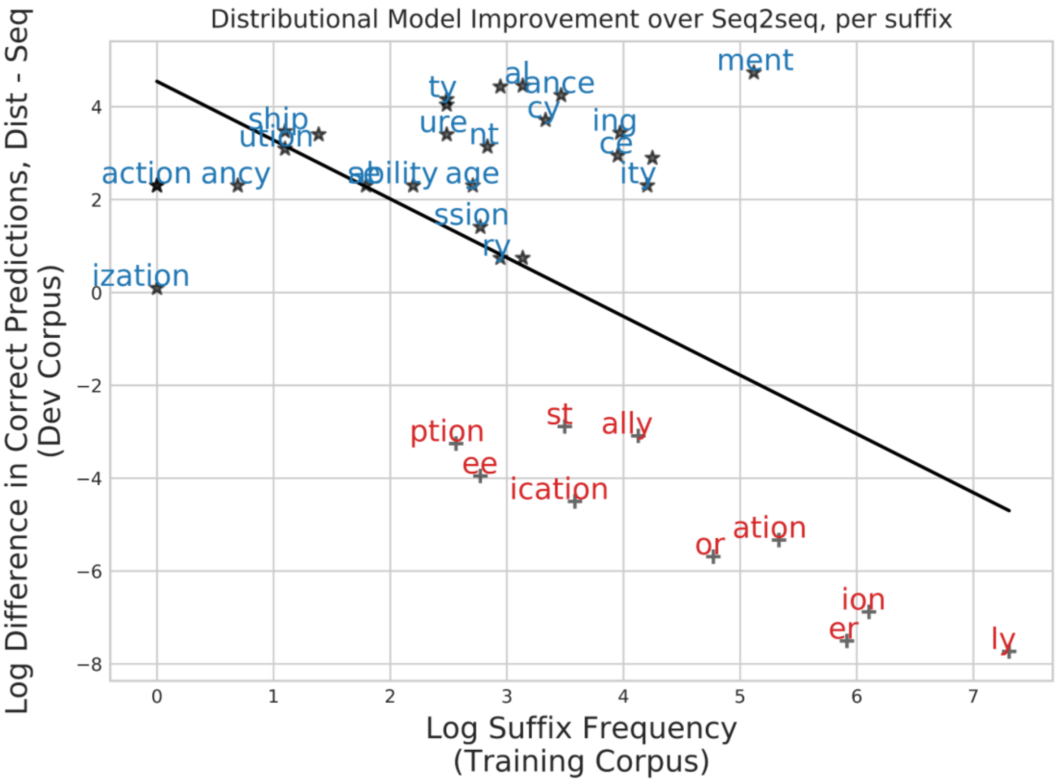
What does each model do well?

Distributional does best



Orthographic does best

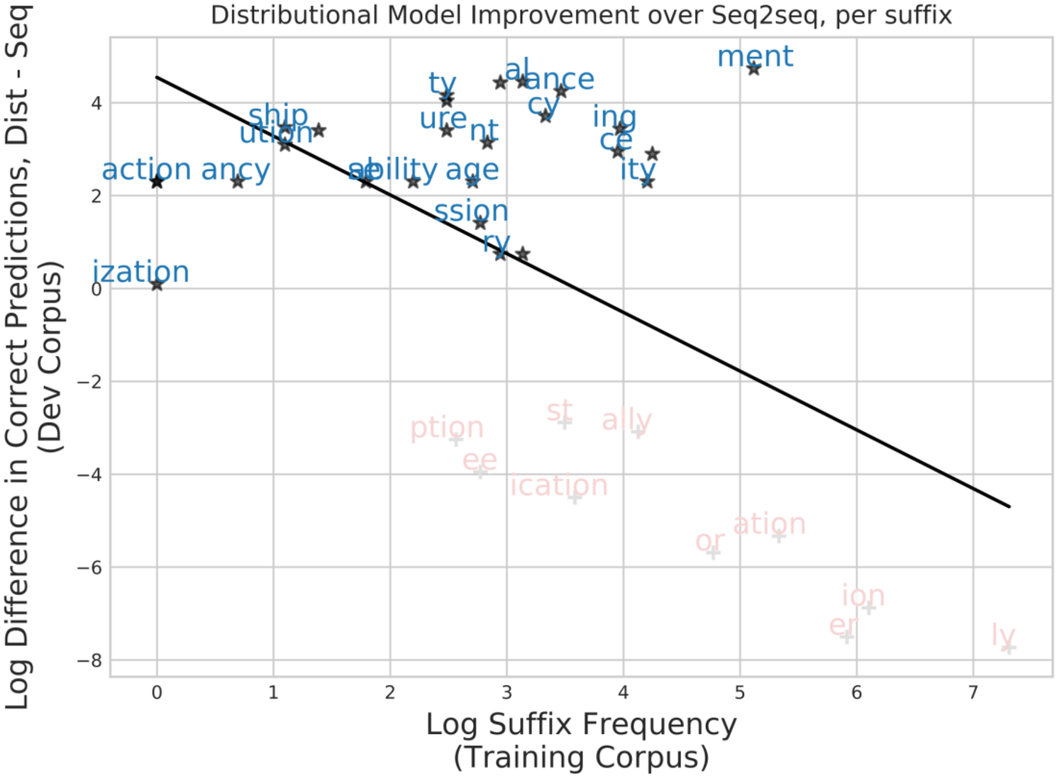




What does each model do well?

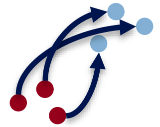
Distributional does best

Orthographic does best

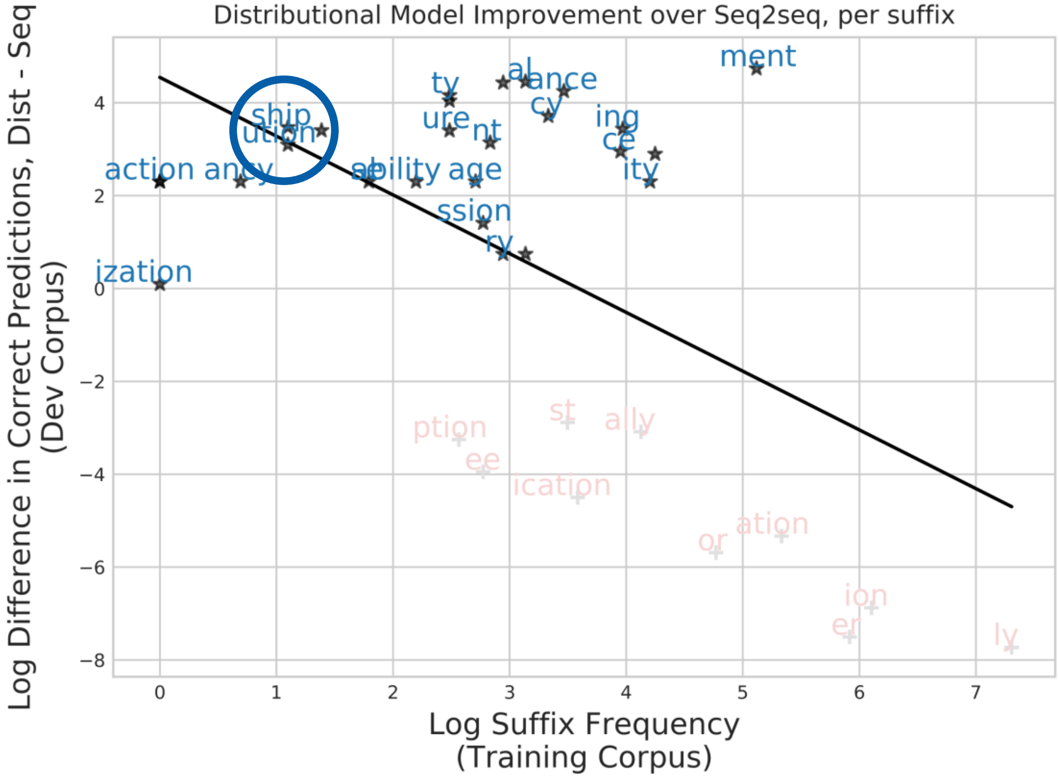


What does each model do well?

Distributional does best



Orthographic does best



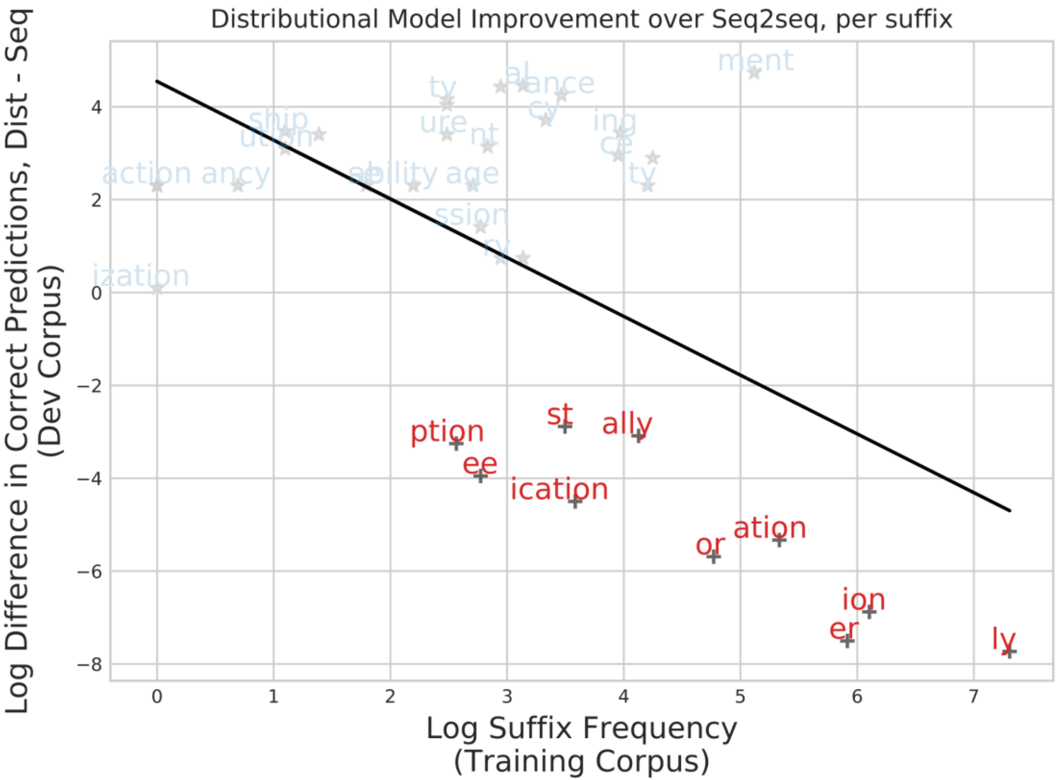
*~~sponsorment~~

sponsor*ship*

What does each model do well?

Distributional does best

Orthographic does best



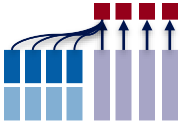
*sponsor*ment*

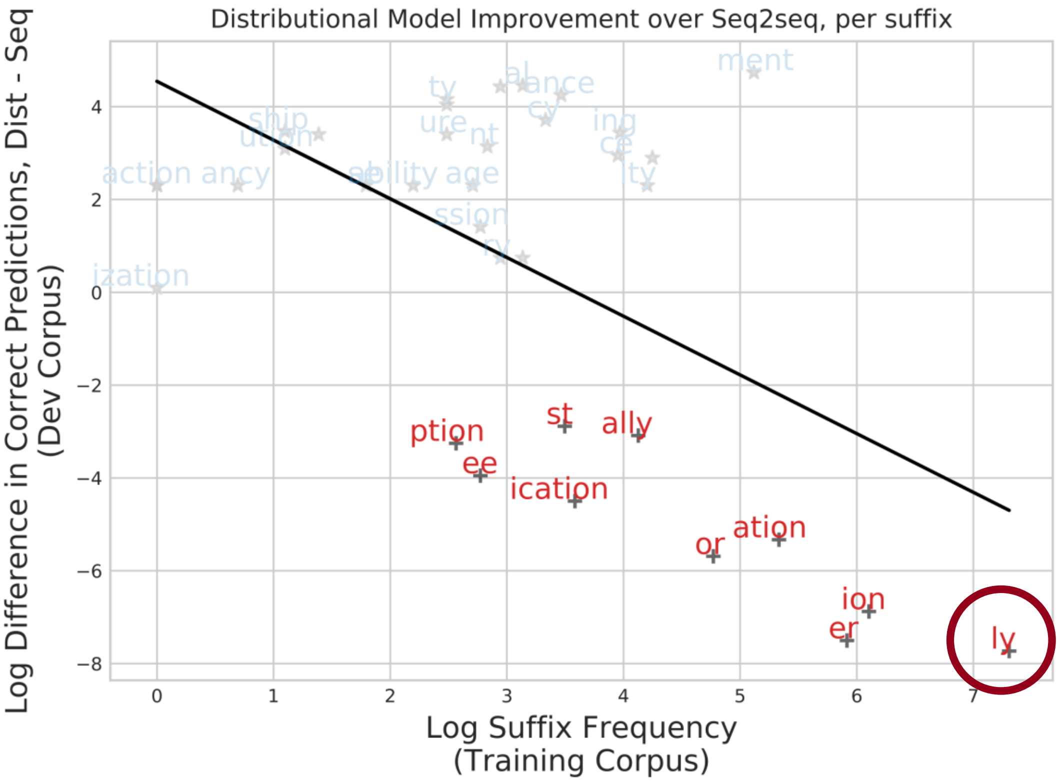
sponsor*ship*

What does each model do well?

Distributional does best

Orthographic does best





*~~sponsorment~~

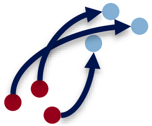
sponsorship

wise $\xrightarrow{\text{ADVERB}}$ wise*ly*

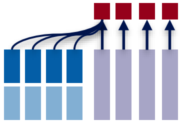
quick $\xrightarrow{\text{ADVERB}}$ quick*ly*

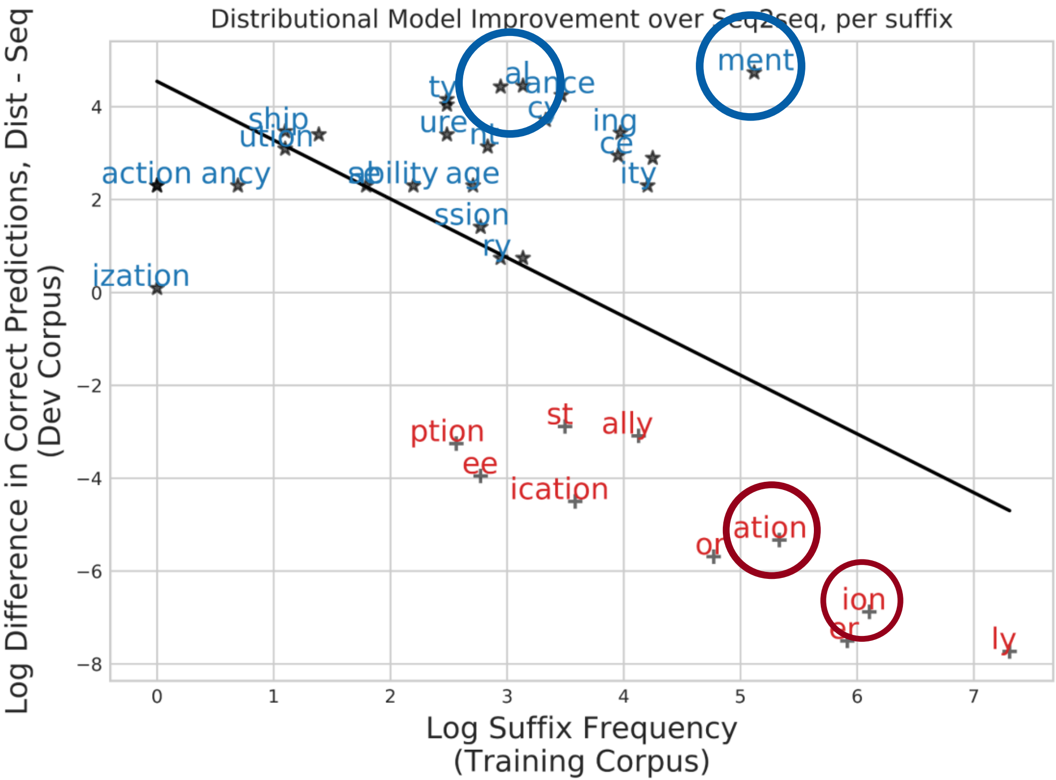
What does each model do well?

Distributional does best



Orthographic does best





*~~sponsorment~~

sponsorship

wise $\xrightarrow{\text{ADVERB}}$ wisely

quick $\xrightarrow{\text{ADVERB}}$ quickly

renew $\xrightarrow{\text{RESULT}}$ renewal

invest $\xrightarrow{\text{RESULT}}$ investment

inspire $\xrightarrow{\text{RESULT}}$ inspiration

Conclusion

- Aggregation model for English derivational morphology
- Dictionary-constrained decoding
- Frequency-based reranking
- Distributional model per-transformation
- Best open- and closed-vocabulary models demonstrate 22% and 37% reduction in error
  - New state-of-the-art results

Code & Data

Code

`https://github.com/danieldeutsch/derivational-morphology`

Data

`https://github.com/ryancotterell/derivational-paradigms`

Powered by dy/net

Penn Engineering

References

- Cotterell et al. 2017, Paradigm completion for derivational morphology. *In EMNLP*

Thank you!