Improving Morphology Induction with Spelling Rules

Jason Naradowsky
University of Massachusetts Amherst
narad@cs.umass.edu

Joint Work with Sharon Goldwater

Outline

- Morphology Induction
- Our Model
- Hyperparameters & Inference
- Experimental Results
- Conclusion

Morphology (Linguistics)

- The study of the internal structure of words:

Antidisestablishmentarianism

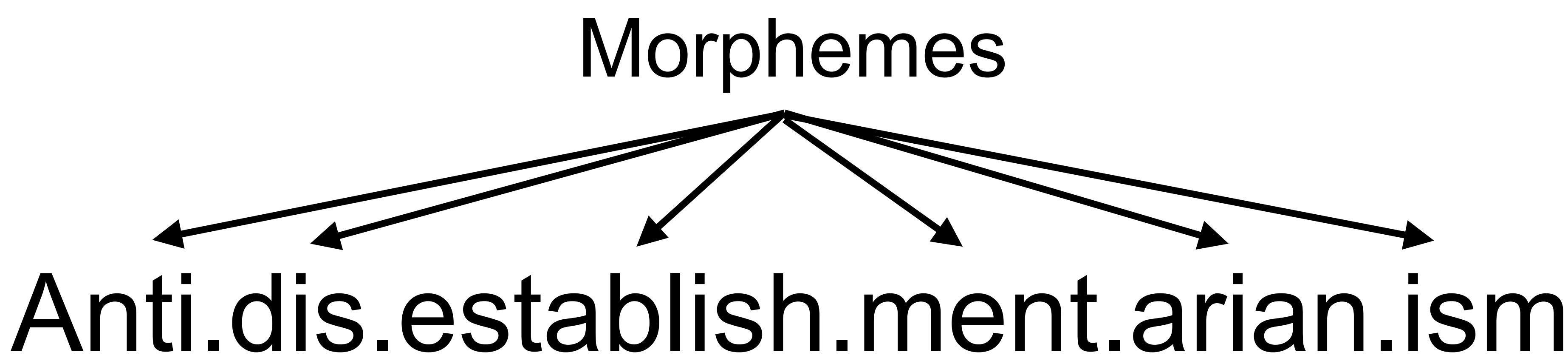
Morphology (Linguistics)

- The study of the internal structure of words:

Anti.dis.establish.ment.arian.ism

Morphology (Linguistics)

- The study of the internal structure of words:



Morphology (Linguistics)

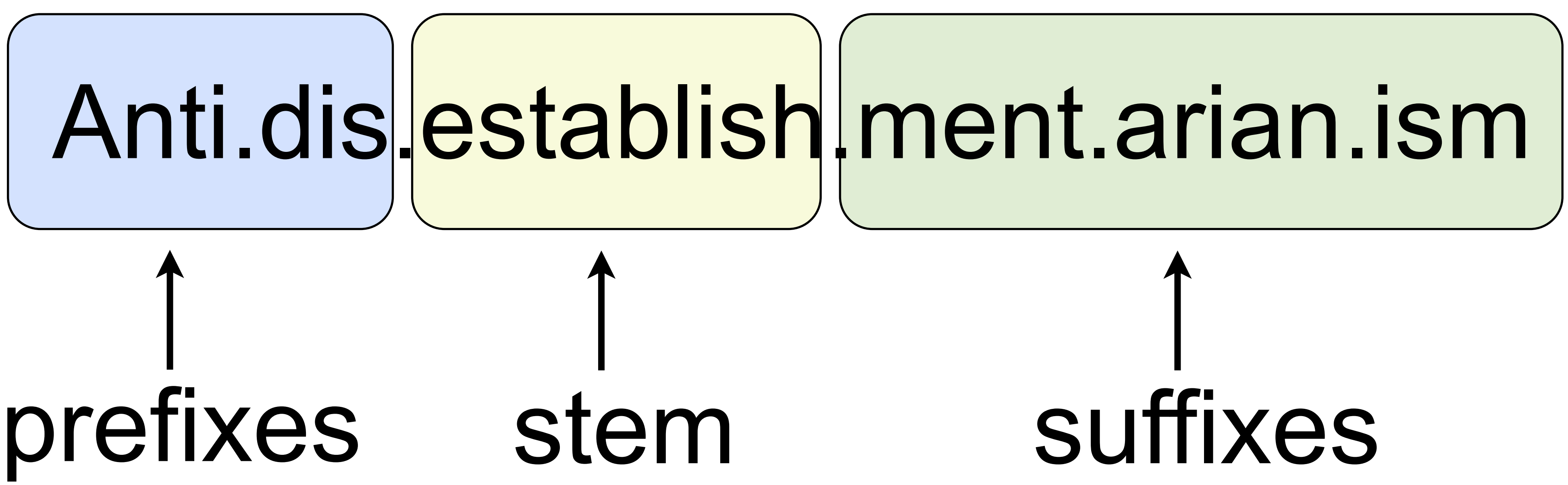
- The study of the internal structure of words:

Anti.dis.establish.ment.arian.ism

↑ stem

Morphology (Linguistics)

- The study of the internal structure of words:



Unsupervised Morphology Induction

- Observing just the words, find the best segmentation:
  - walking → walk.ing

- Applications:
  - Important component in many NLP tasks
  - Especially useful for morphologically-rich languages (Finnish, Arabic, Hebrew)
  - Cognitive Science: How do children learn this?

Underlying Assumption:

- User's Goal: Find best (linguistic) solution.
- System Goal: Find most concise solution.

Too Many Stems	Too Many Suffixes	Just Right
walk.	wa.lk	walk.
walks.	wa.lks	walk.s
walking.	wa.lking.	walk.ing
talk.	ta.lk	talk.
talking.	ta.lking	talk.ing
cat.	cat.	cat.
cat.s	cat.s	cat.s
Morphs: 6+2=8	3+5=8	3+3=6

Underlying Assumption:

- User's Goal: Find best (linguistic) solution.
- System Goal: Find most concise solution.

Too Many Stems	Too Many Suffixes	Just Right
walk.	wa.lk	walk.
walks.	wa.lks	walk.s
walking.	wa.lking.	walk.ing
talk.	ta.lk	talk.
talking.	ta.lking	talk.ing
cat.	cat.	cat.
cat.s	cat.s	cat.s
Morphs: 6+2=8	3+5=8	3+3=6

Underlying Assumption:

- User's Goal: Find best (linguistic) solution.
- System Goal: Find most concise solution.

Too Many Stems	Too Many Suffixes	Just Right
walk.	wa.lk	walk.
walks.	wa.lks	walk.s
walking.	wa.lking.	walk.ing
talk.	ta.lk	talk.
talking.	ta.lking	talk.ing
cat.	cat.	cat.
cat.s	cat.s	cat.s
Morphs: 6+2=8	3+5=8	3+3=6

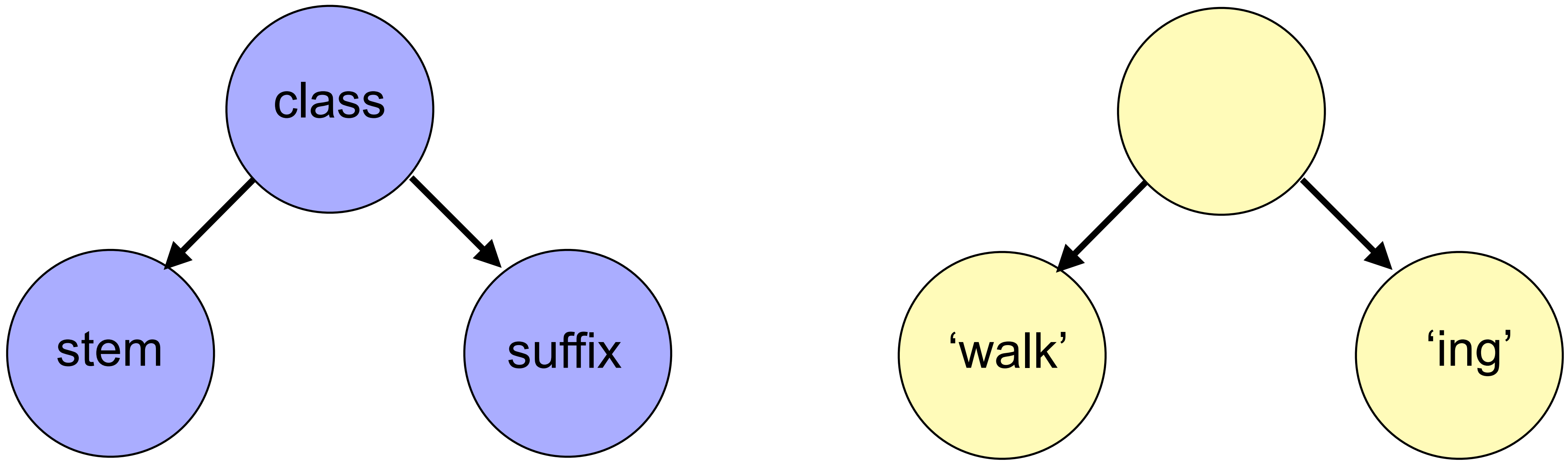
Bayesian Morphology Induction

(Goldwater 2006)

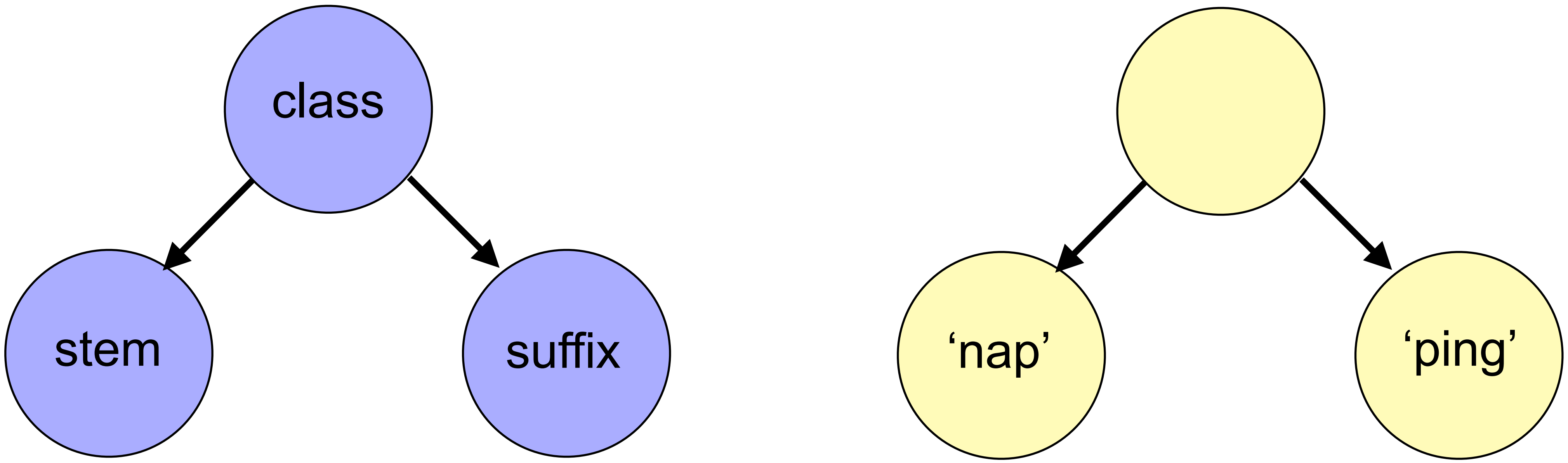
P(word) = P(class, stem, suffix) =
 P(class) x
 P(stem | class) x
 P(suffix | class)

- Each word consists of a stem and a suffix
  - (suffix can be the empty string)
- Multinomials with symmetric Dirichlet priors
  - No bias means most concise solution preferable

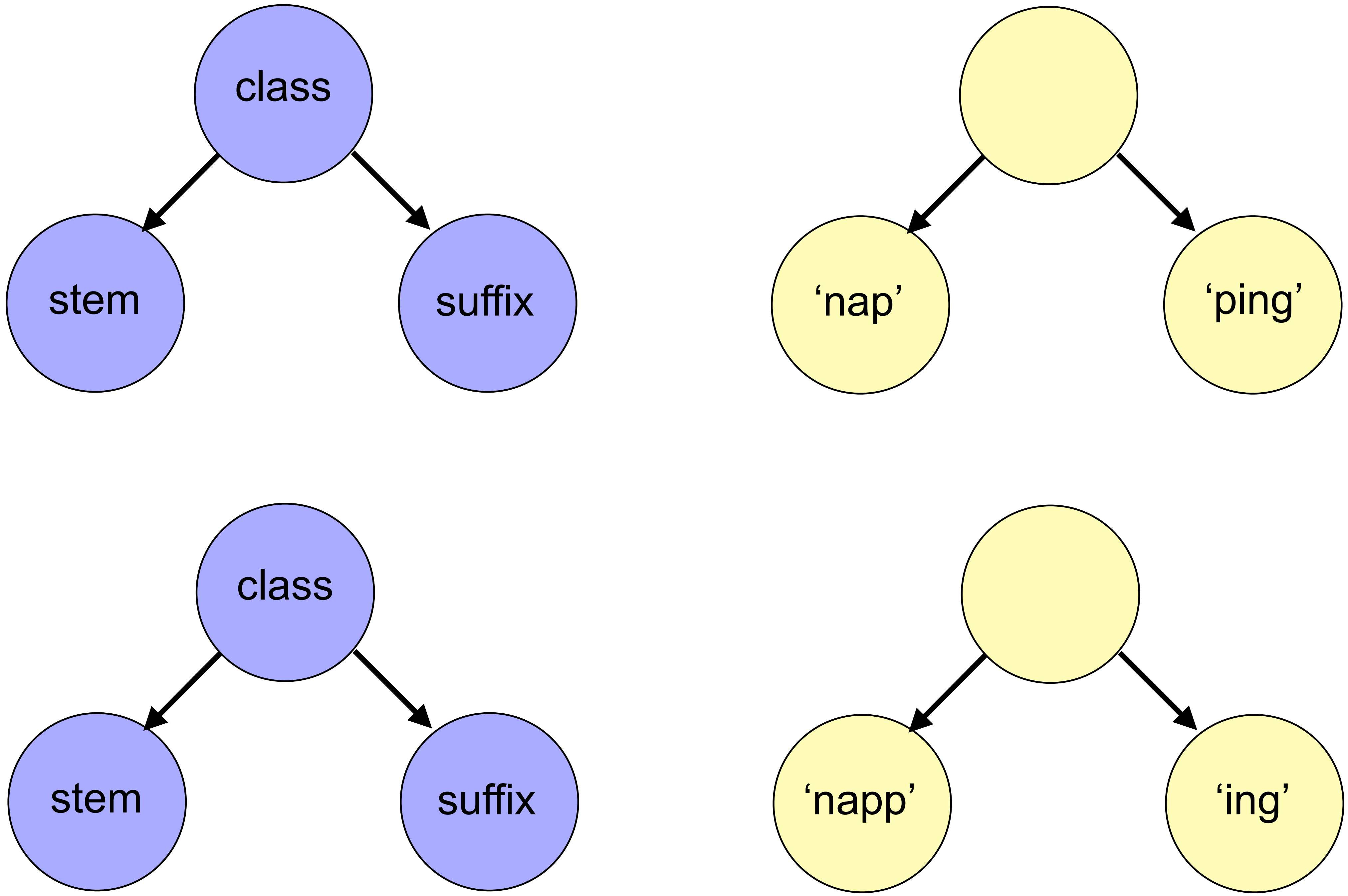
Generative Process: ‘walking’



Generative Process??: 'napping'



Generative Process??: ‘napping’



Spelling Rules

$$\varepsilon \rightarrow p \;/\; ap \,\_\, i$$

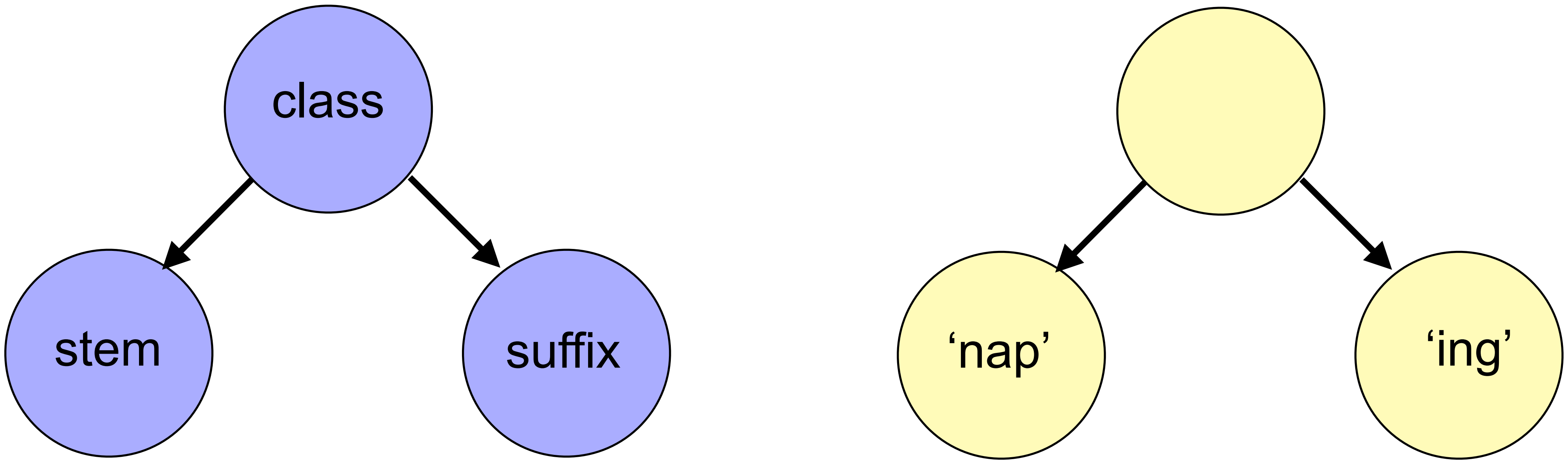
- ε: original character
- p: transform character
- ap: left context
- i: right context

- Rules capture a one-character transformation in a particular context.
- 3 Types: Insertions, Deletions, and Null (no transformation)
- Left context more important in English
  - (we find 2 character left contexts most useful)

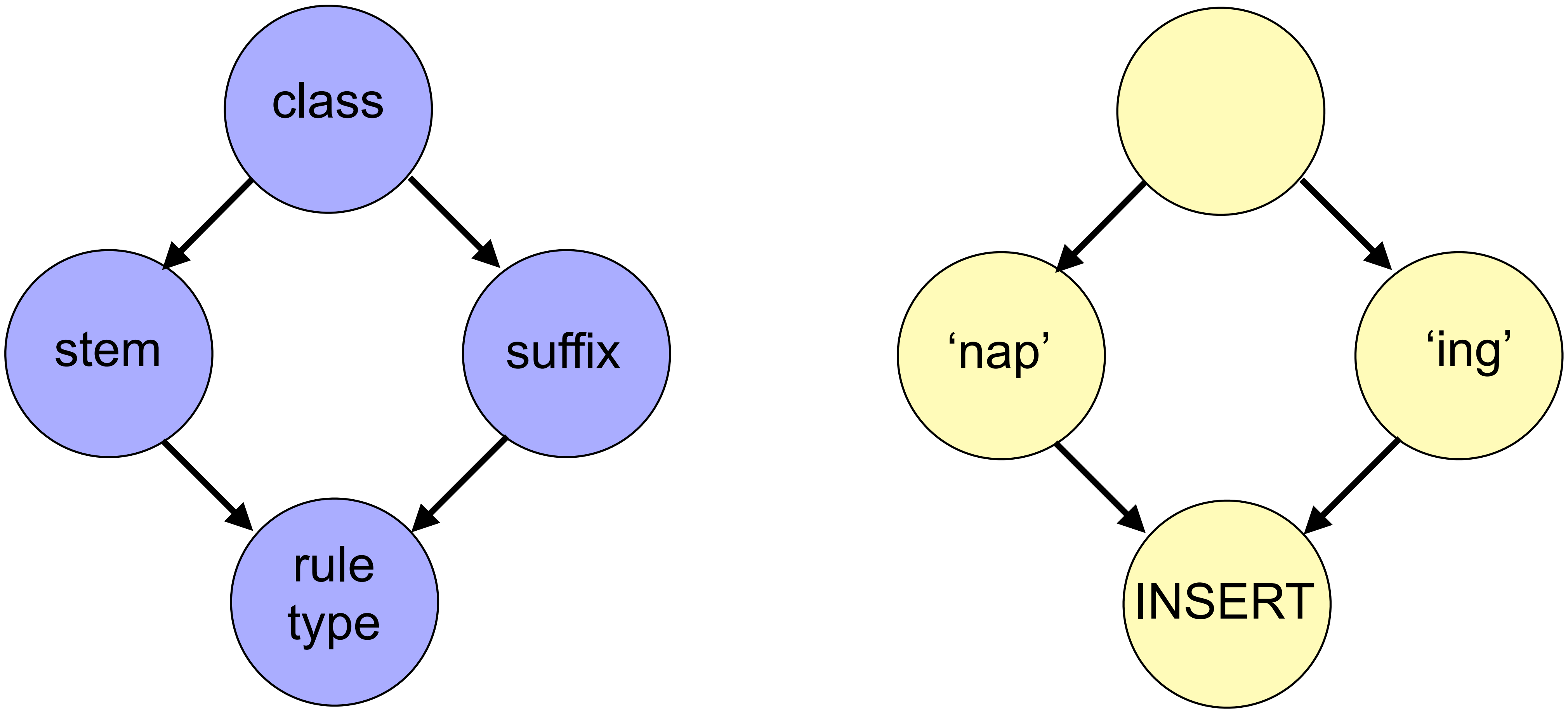
Outline

- Morphology Induction
- Our Model
- Hyperparameters & Inference
- Experimental Results
- Conclusion

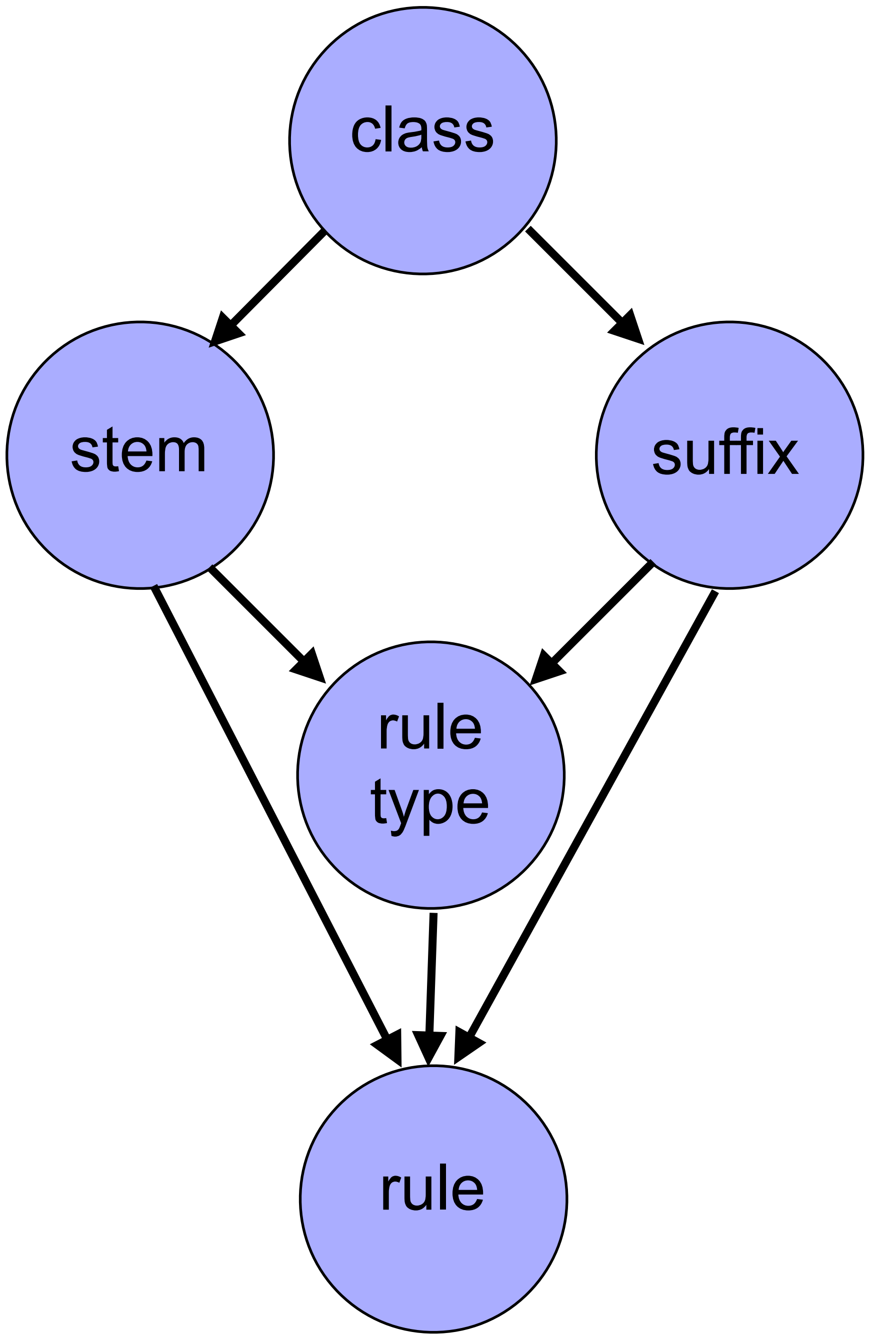
A New Generative Process:

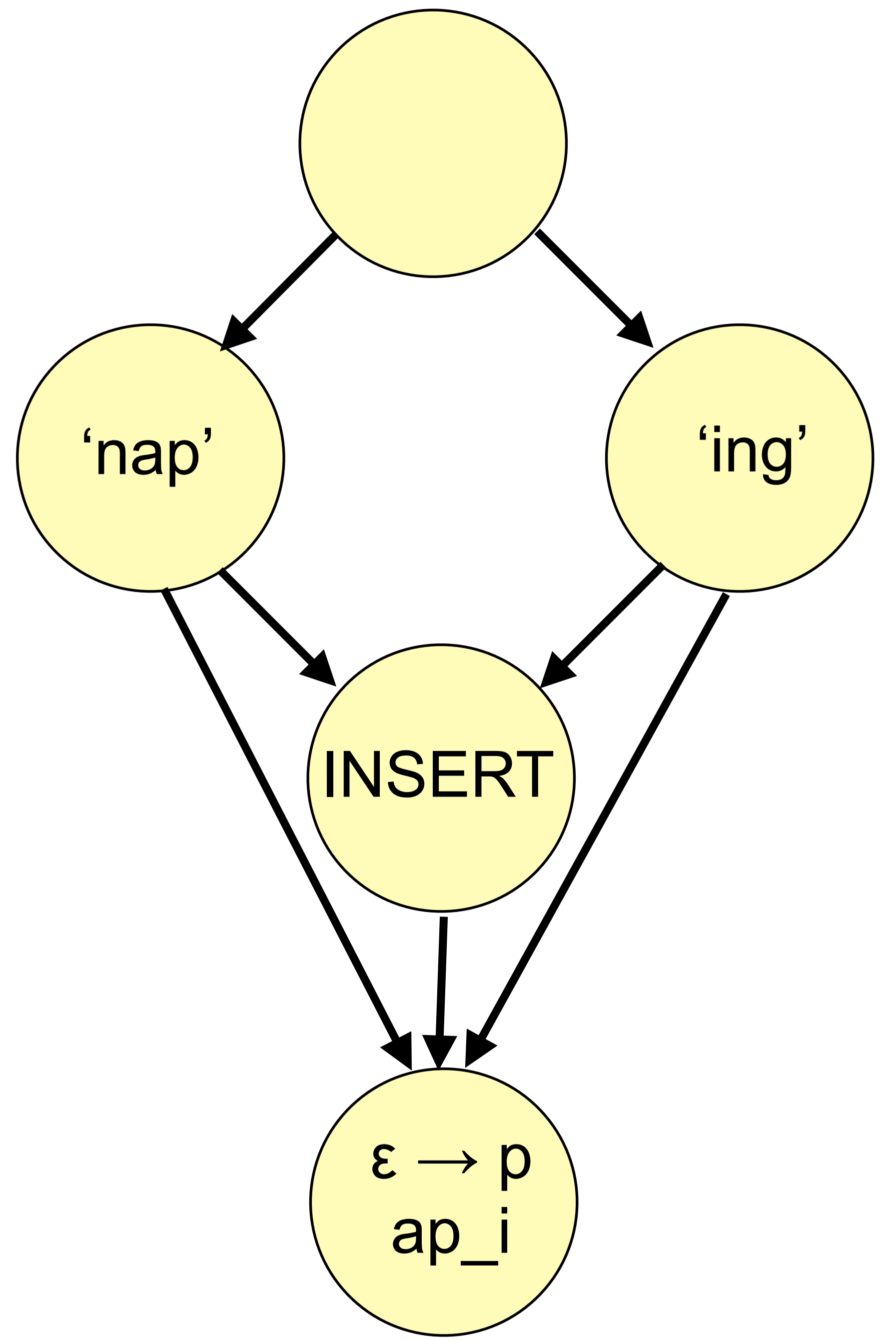


A New Generative Process:



A New Generative Process:





Our Model

P(class, stem, suffix, rule type, rule) =

P(class) x

P(stem | class) x

P(suffix | class) x

P(rule type | context(stem, suffix)) x

P(rule | rule type, context(stem, suffix))

rule type $\in$ { Insertion, Deletion, Null }

- Greatly increases search space:
  - About 28 times more possible solutions per word!

Outline

- Morphology Induction
- Our Model
- Hyperparameters & Inference
- Experimental Results
- Conclusion

Inference

- Alternate between:
  - Gibbs Sampling for the latent variables
    - (class, stems, suffix, etc)
  - Hyperparameter Updates
    - (update hyperparameters over priors on variables)
    - minimize free parameters!
- We run for 5 epochs of:
  - 10 Gibbs Sampling Iterations
  - 10 hyperparameter iterations
- Convergence much earlier

Hyperparameters

- Induced for class, stem, suffix, and rule variables
- Learn hyperparameters using Minka's fixed-point method (Minka, 2003)
- Inducing all is principled, but also a computational burden
- Rule type prior set by linguistic intuition:
  - hyp(INSERTION) = .001
  - hyp(DELETION) = .001
  - hyp(NULL) = .5

Outline

- Morphology Induction
- Our Model
- Hyperparameters & Inference
- Experimental Results
- Conclusion

Data Sets & Evaluation

- 7487 different verbs from Wall Street Journal
- Gold Standard: CELEX lexical database
  - surface segmentation: walk.ing
  - abstract representation: 50655+pe

- Evaluation Metrics:
  - Underlying form accuracy
  - Pairwise precision and recall

Underlying Form Accuracy

- Construct the underlying stem from derivational data contained in the CELEX (using lemma ID number)
- Lookup suffix in dictionary:
  - e3S : -s
  - a1S : -ed
  - pe : -ing
- Match strings - UFA is % correct

Pairwise Precision and Recall

Word	Found	Gold
state	state+ε ε → ε	44380+i
stating	state+ing e → ε	44380+pe
states	stat.es ε → ε	44380+a1S
station	stat+ion ε → ε	44405+i

Pairwise Precision and Recall

Word	Found	Gold
state	state+ε ε → ε	44380+i
stating	state+ing e → ε	44380+pe
states	stat.es ε → ε	44380+a1S
station	stat+ion ε → ε	44405+i

Pairwise Precision and Recall

Word	Found		Gold
state	state+ε	ε → ε	44380+i
stating	state+ing	e → ε	44380+pe
states	stat.es	ε → ε	44380+a1S
station	stat+ion	ε → ε	44405+i

Pairwise Precision and Recall

Word	Found		Gold
state	state+ε	ε → ε	44380+i
stating	state+ing	e → ε	44380+pe
states	stat.es	ε → ε	44380+a1S
station	stat+ion	ε → ε	44405+i

1 match out of 1 arcs = 100% PP for this stem

Pairwise Precision and Recall

Word	Found	Gold
state	state+ε ε → ε	44380+i
stating	state+ing e → ε	44380+pe
states	stat.es ε → ε	44380+a1S
station	stat+ion ε → ε	44405+i

Pairwise Precision and Recall

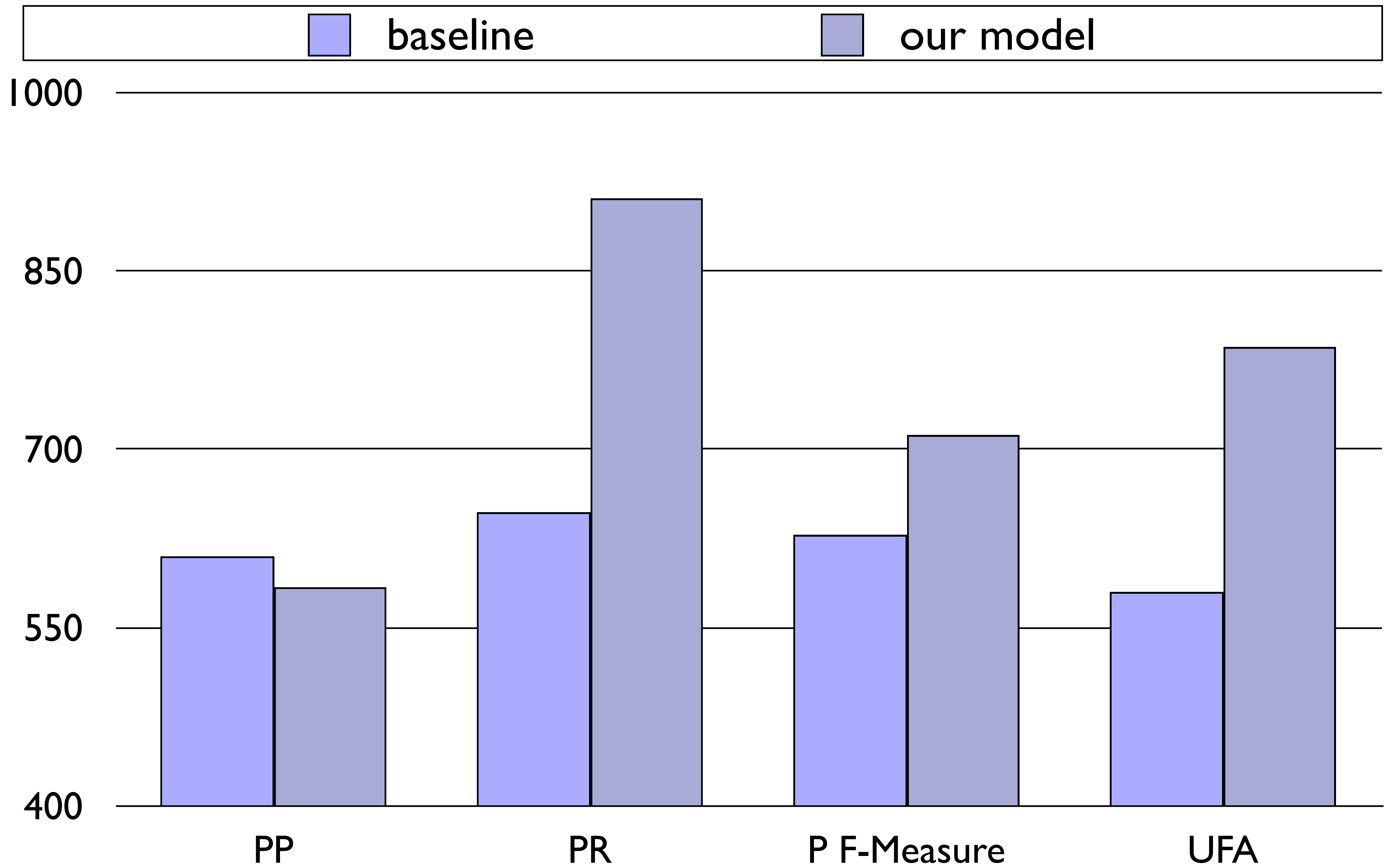
Word	Found	Gold
state	state+ε ε → ε	44380+i
stating	state+ing e → ε	44380+pe
states	stat.es ε → ε	44380+a1S
station	stat+ion ε → ε	44405+i

Pairwise Precision and Recall

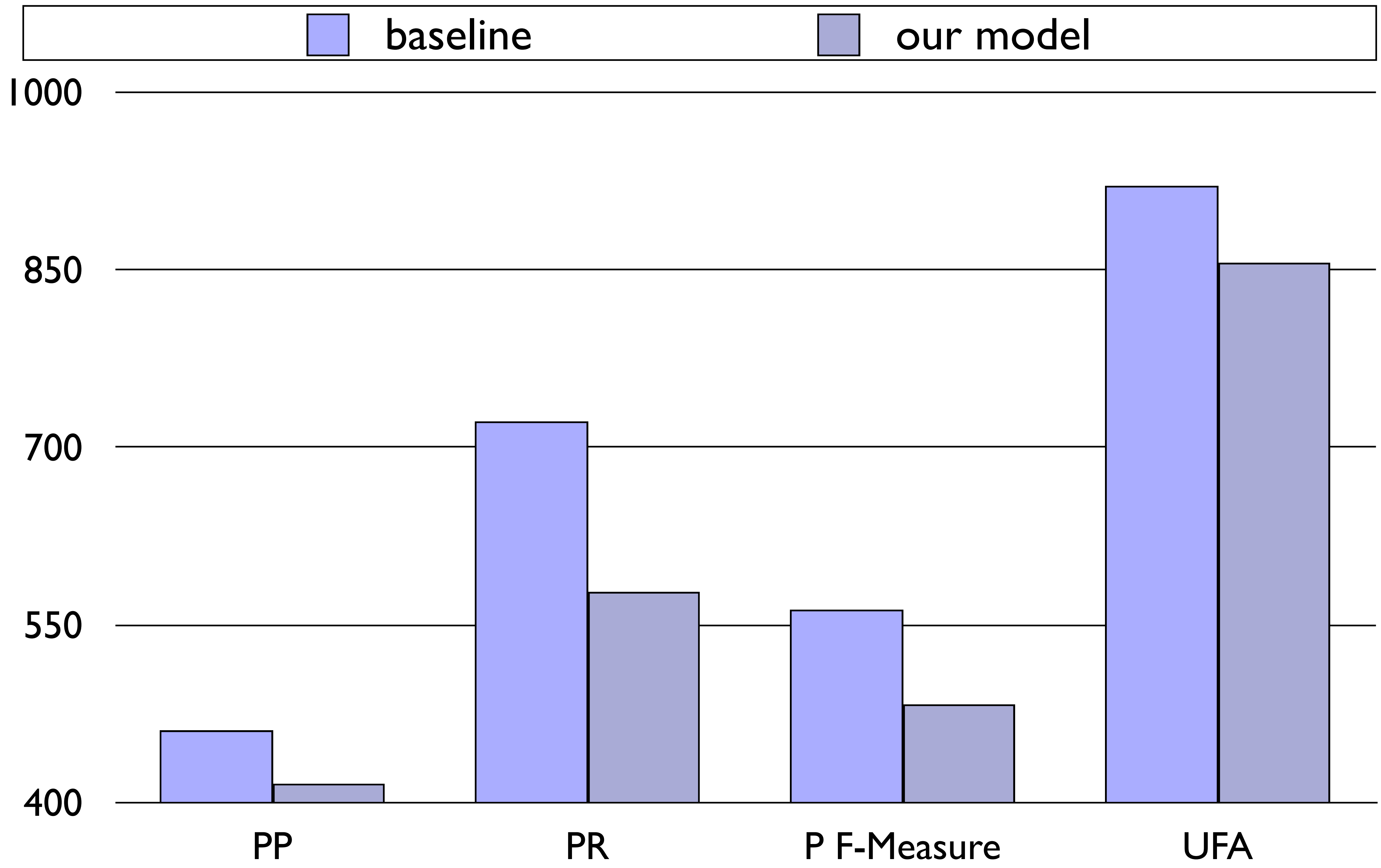
Word	Found	Gold
state	state+ε ε → ε	44380+i
stating	state+ing e → ε	44380+pe
states	stat.es ε → ε	44380+a1S
station	stat+ion ε → ε	44405+i

1 correct arc out of 2 arcs = %50 Recall for this stem

Results: Stems



Results: Suffixes



Induced Rules:

Freq	Rule	Example
468	e → ε when before i	abate, abating
41	ε → e when after sh/ss/ch	match, matches
29	ε → p after p, before i or e	nap, napping

Of the top 20 types of induced rules,
568 of 623 correct = 91 %

Incorrect rules: fated explained as fates.d with s-deletion
rates explained as rat.s with an e-insertion

Conclusions

- Orthographic rules can help in morphology induction
- Greatly increases search space
- Joint inference over complimentary tasks can overcome the search burden and significantly improve performance in particular parts of task
- This may allow unsupervised generative models to compete more closely with unsupervised discriminative models (with contrastive estimation)

Future Work

- Extend to multiple suffixes
  - Test on more representative language samples
  - Test on more languages
- Leverage phonological information for asymmetric priors
  - Once we know ‘p’ is often doubled, and ‘t’ is similar to ‘p’, should imply ‘t’ may also often be doubled
  - May allow for character-to-character transformations
- Hierarchical Models
  - More like grammar induction than segmentation
  - Capture interaction between prefixes and suffixes