Unsupervised Multilingual Learning for POS Tagging

Benjamin Snyder, Tahira Naseem
Jacob Eisenstein and Regina Barzilay
MIT

Unsupervised Learning in NLP

Unsupervised Learning in NLP

- Has focused on monolingual settings

Unsupervised Learning in NLP

- Has focused on monolingual settings
- Performance still lags supervised learning

Unsupervised Learning in NLP

- Has focused on monolingual settings
- Performance still lags supervised learning

Question: can we improve *monolingual* performance when *multilingual* parallel data is available at training time?

Multilingual Learning for POS Tagging

Input:

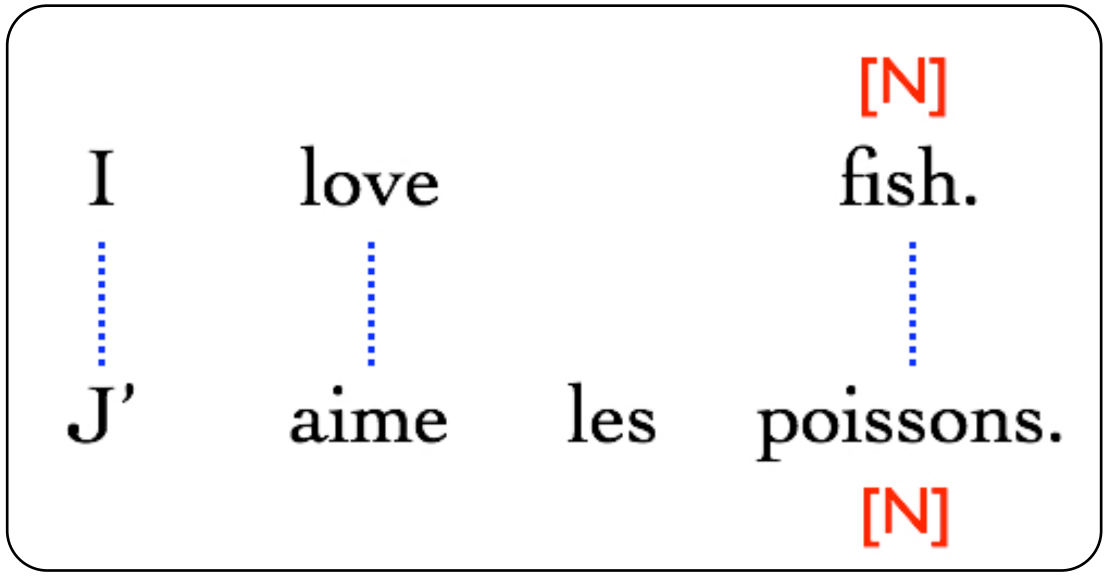
untagged bilingual parallel corpus

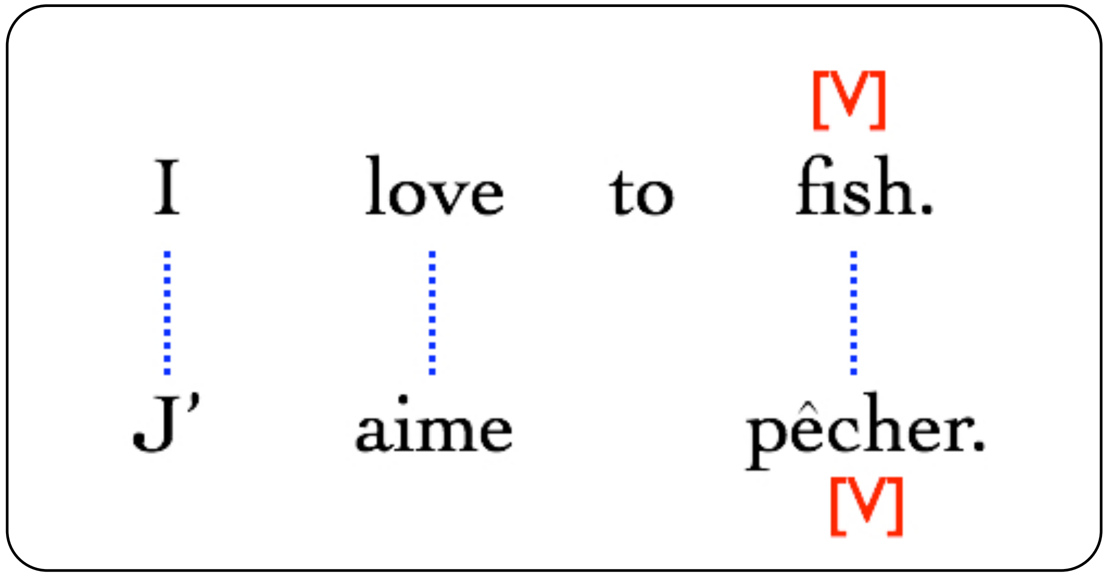
Goal:

Induce a POS tagger for each language
(test on monolingual data)

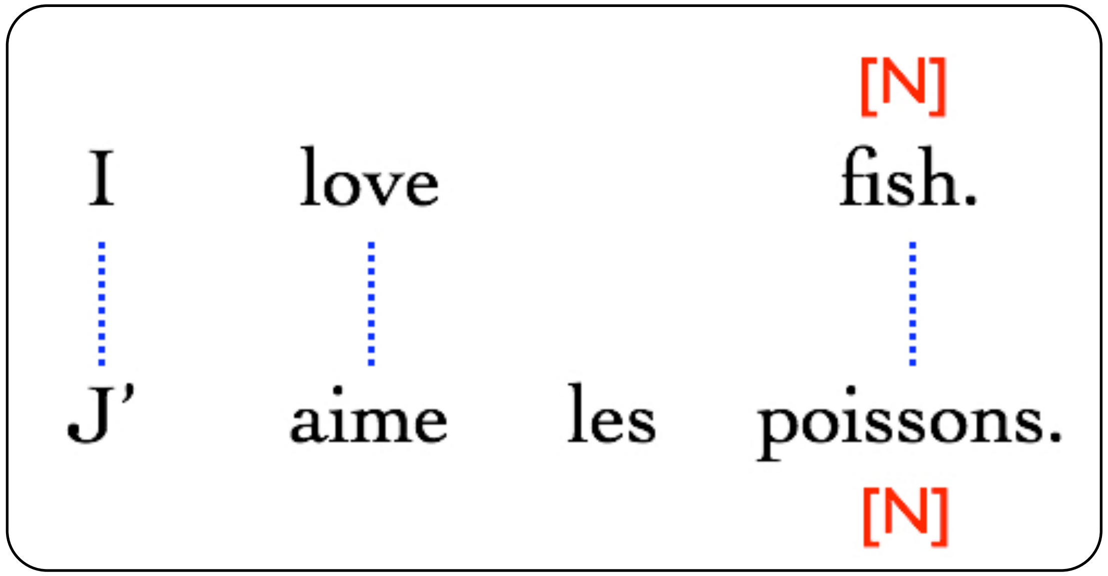
[P]	[V]		[N]
I	love		fish.
J'	aime	les	poissons.
[P]	[V]	[D]	[N]

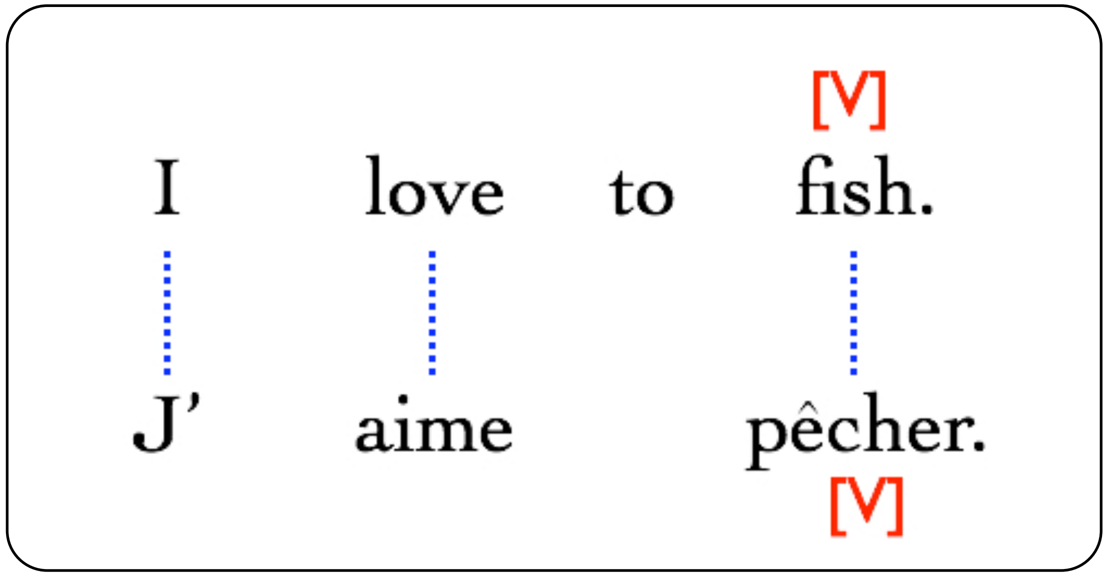
Motivation for Multilingual Learning





Motivation for Multilingual Learning

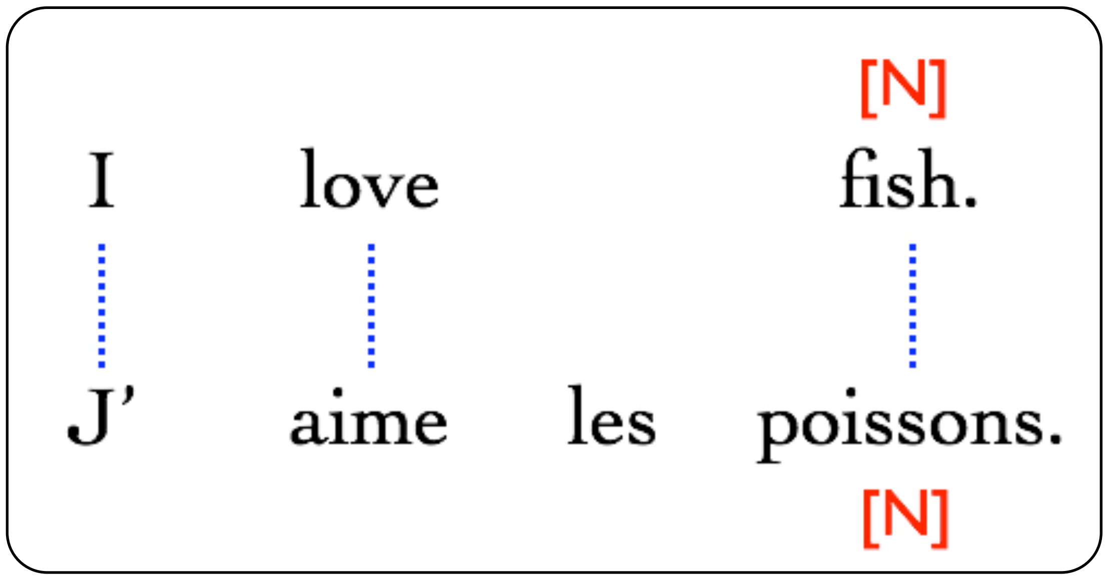


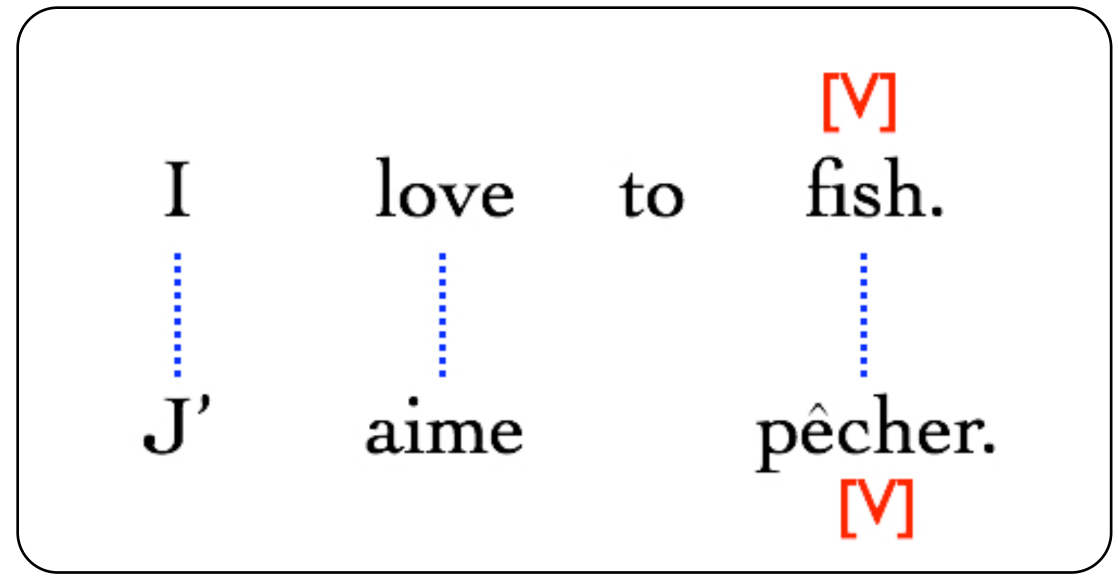


- Learn from differences in lexical ambiguity

  fish/poissons [N] vs. fish/pêcher [V]

Motivation for Multilingual Learning





- Learn from differences in lexical ambiguity

  fish/poissons [N] vs. fish/pêcher [V]

- Learn from differences in structural ambiguity

  (1) determiner "*les*" signals noun

  (2) "*to*" signals infinitival verb

Related Work

- Projection (Yarowsky & Ngai 2001, Feldman et al 2006)
  - Supervised data available in source language
  - Goal: transfer annotations to target language

- Synchronous grammars for MT
  (Wu & Wong 1998, Chiang 2005)

Bilingual Graphical Models

Desiderata:

Symmetric model:

- No supervision on either side
- Information flows both ways

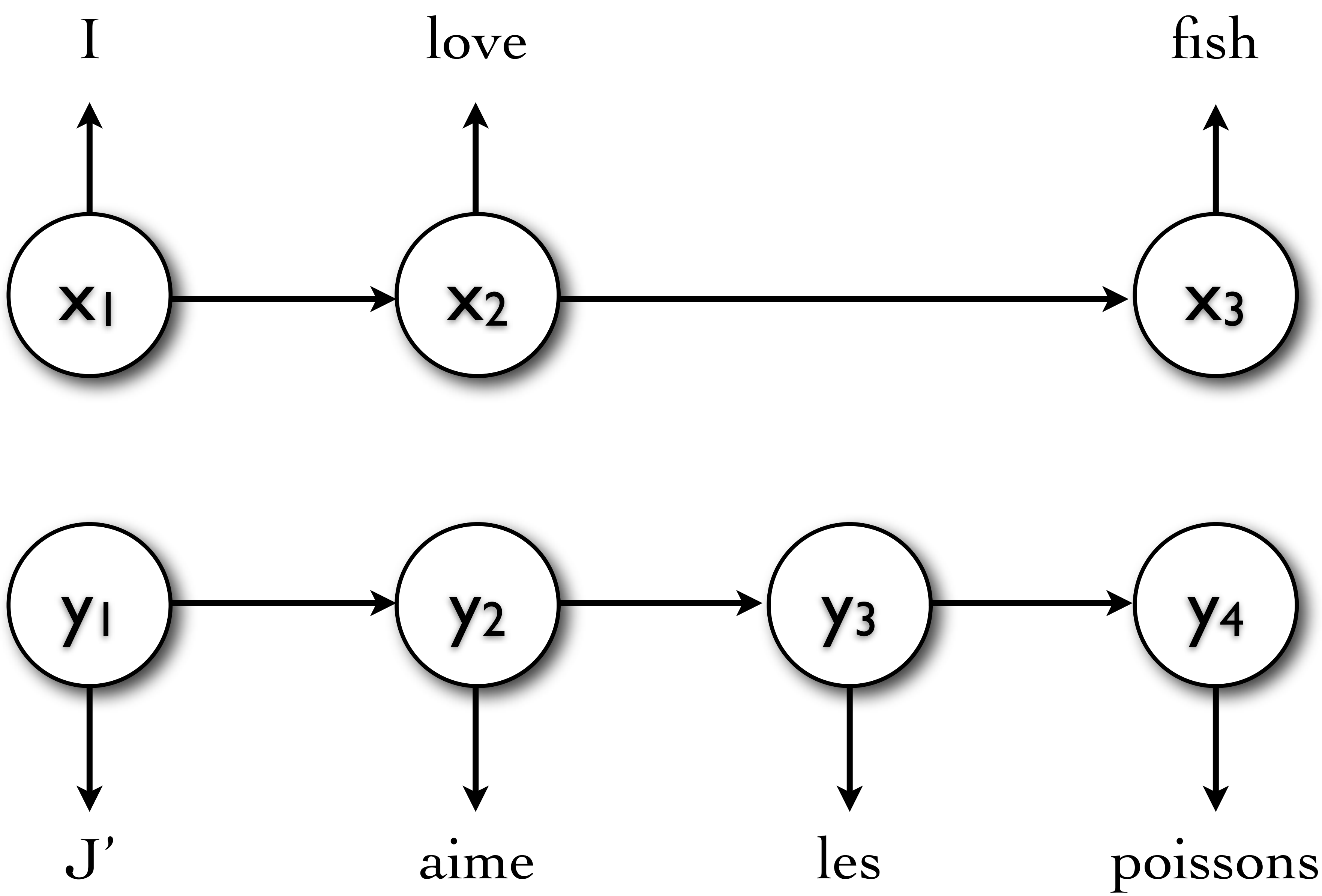
Minimalist approach:

- Allow language specific idiosyncrasies

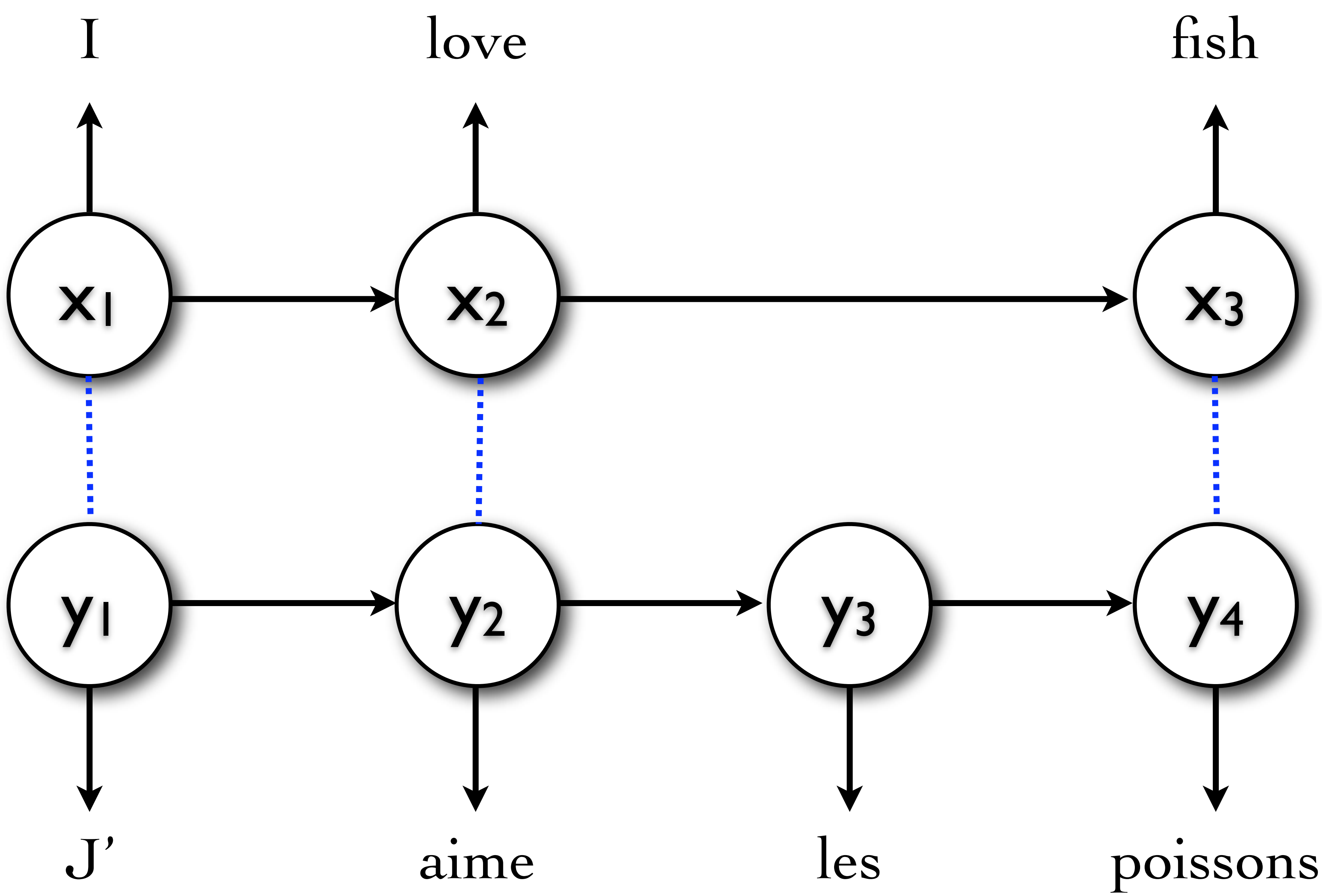
  different sentence lengths, tags, *tagsets* etc

- Avoid over-parameterization

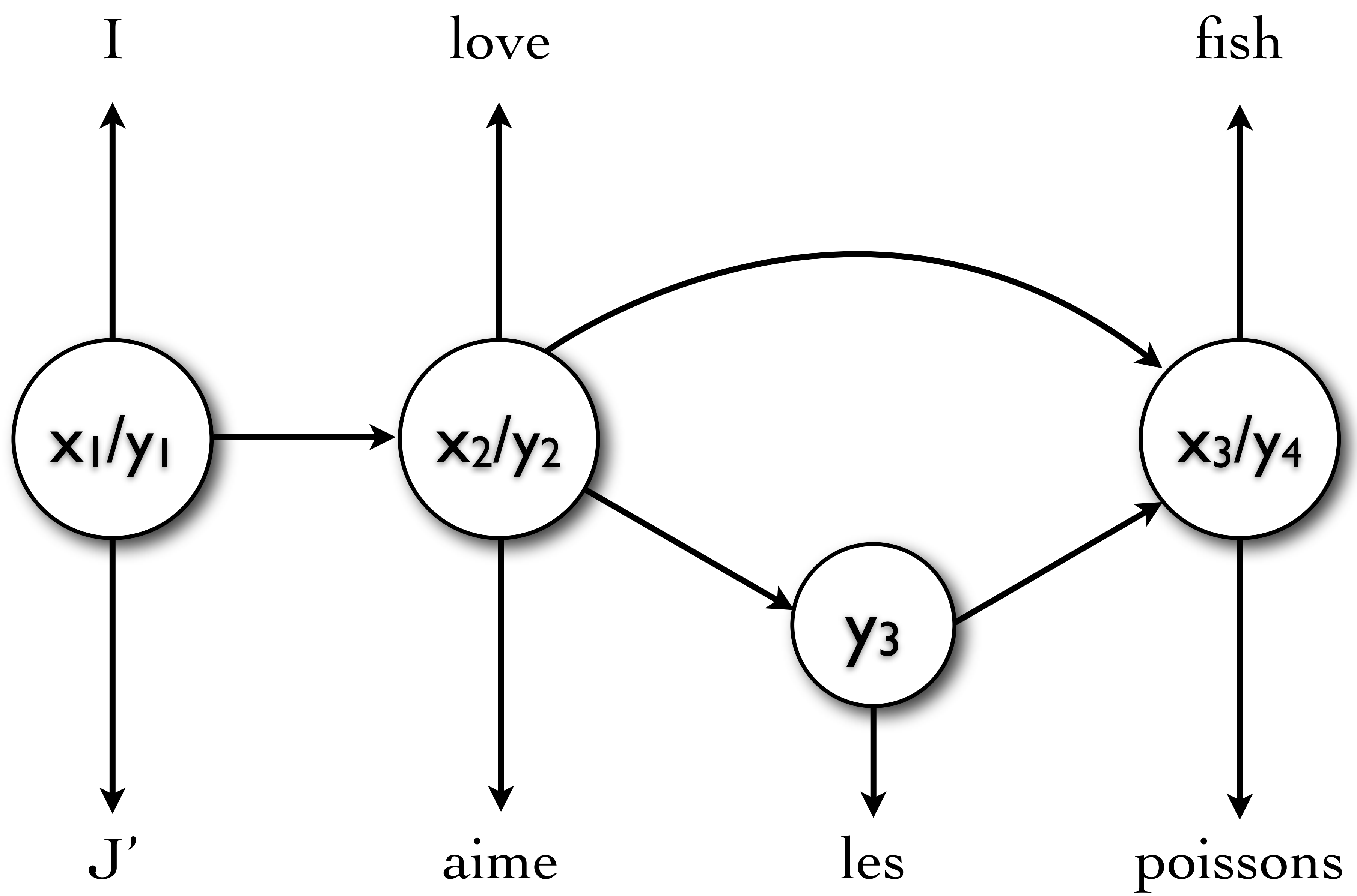
(1) Two Monolingual HMM's



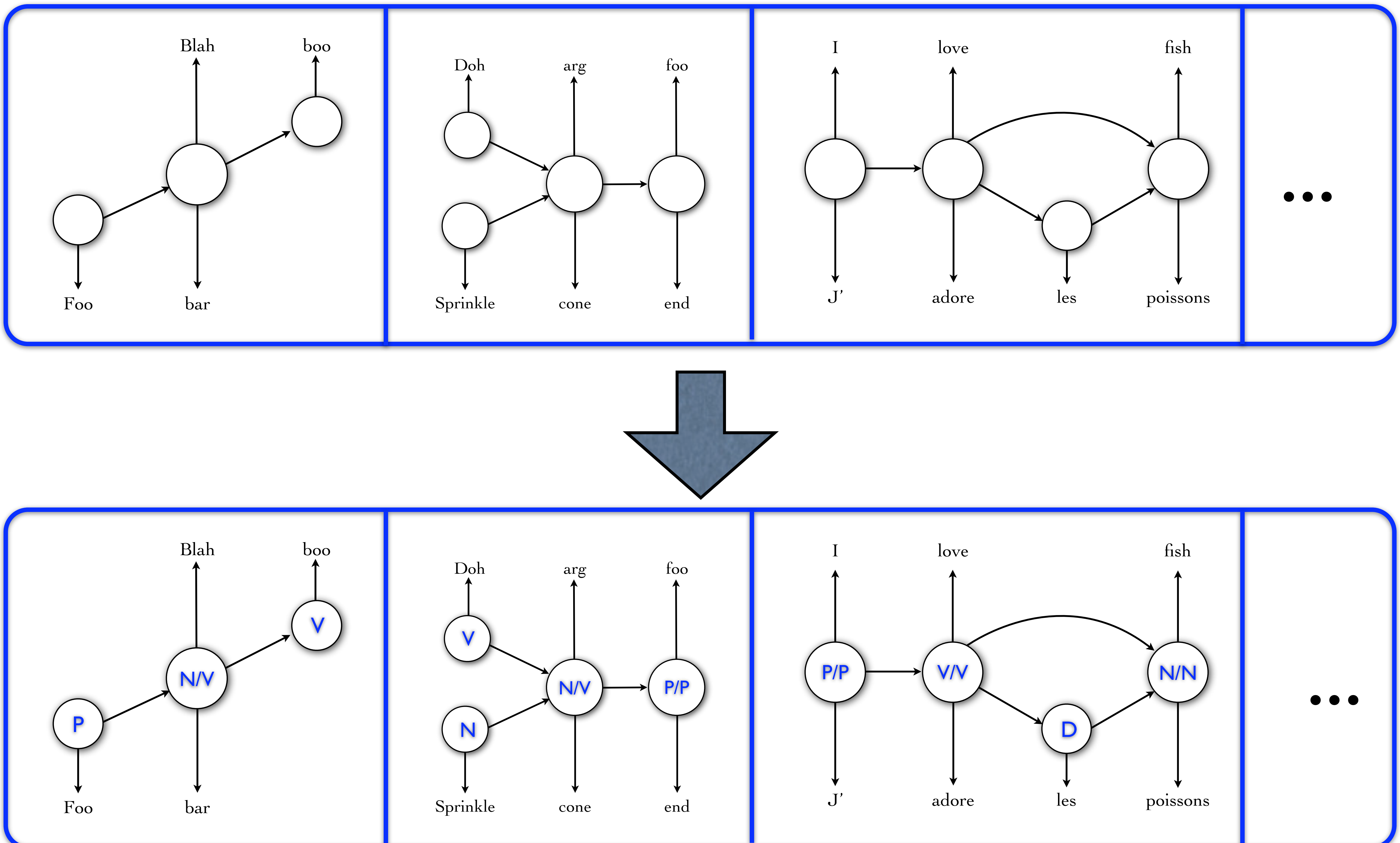
(2) Get Alignments (using GIZA++)



(3) Form Bilingual Model

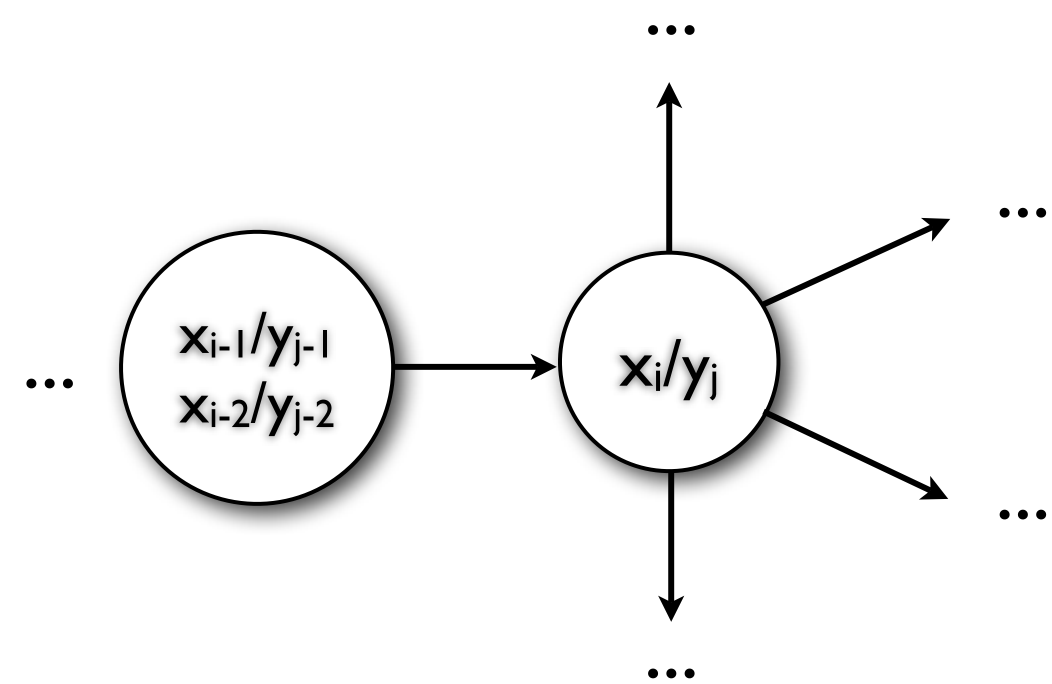


Learning Task

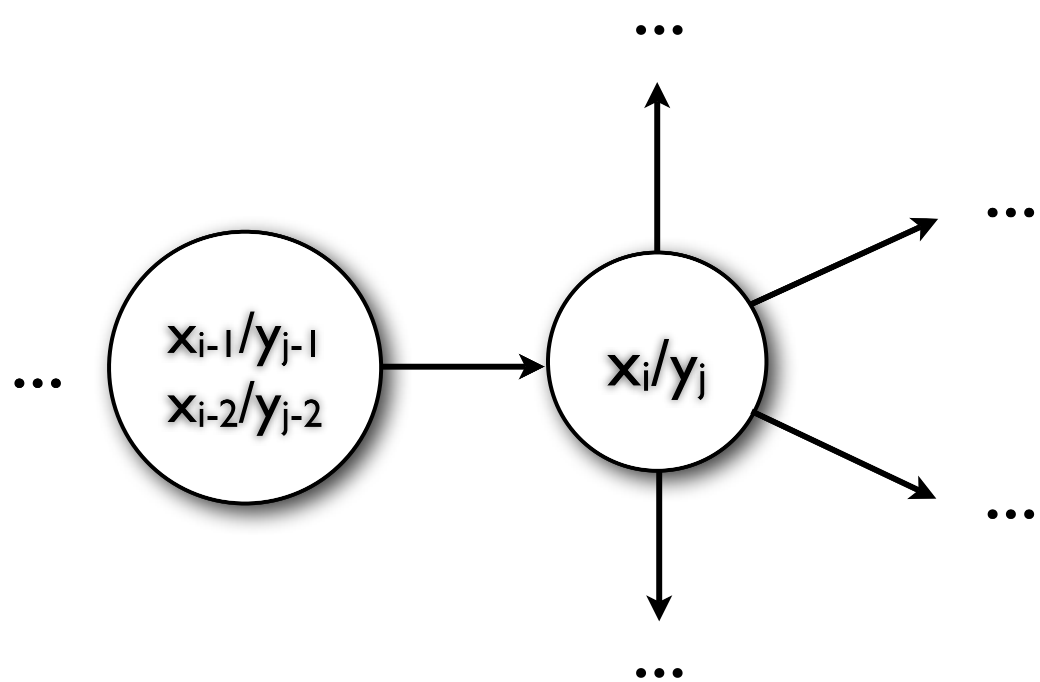


How to Parameterize

How to Parameterize

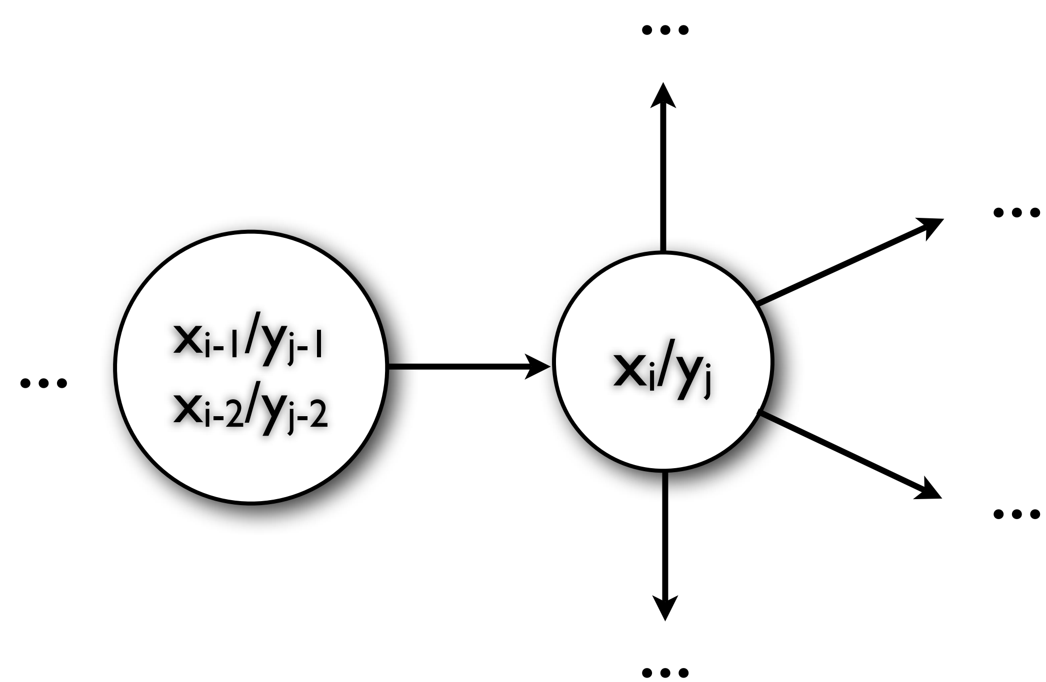


How to Parameterize



<u>Naive parameterization</u>: multinomial over merged tag pair, conditioned on both languages' previous tags.

How to Parameterize



Naive parameterization: multinomial over merged tag pair, conditioned on both languages' previous tags.

- No parameter sharing
- For trigram tagger with 13 tags:

  28,561 unrelated multinomials (13^4)
  each of dimension 169 (13^2)

Instead, we define the generative probability of merged tag pair (x_i, y_j) in terms of three factors:

$$P(x_i, y_j | x_{i-1}, x_{i-2}, y_{i-1}, y_{i-2}) \propto$$
$$P(x_i | x_{i-1}, x_{i-2}) P(y_j | y_{j-1}, y_{j-2}) P(x_i, y_j)$$

Instead, we define the generative probability of merged tag pair (x_i, y_j) in terms of three factors:

$$P(x_i, y_j | x_{i-1}, x_{i-2}, y_{i-1}, y_{i-2}) \propto$$
$$P(x_i | x_{i-1}, x_{i-2}) P(y_j | y_{j-1}, y_{j-2}) P(x_i, y_j)$$

Transition probability in each language

Instead, we define the generative probability of merged tag pair (x_i, y_j) in terms of three factors:

$$P(x_i, y_j | x_{i-1}, x_{i-2}, y_{i-1}, y_{i-2}) \propto$$
$$P(x_i | x_{i-1}, x_{i-2}) P(y_j | y_{j-1}, y_{j-2}) P(x_i, y_j)$$

Transition probability in each language

"Coupling" probability: compatibility of tag pair

Instead, we define the generative probability of merged tag pair (x_i, y_j) in terms of three factors:

$$P(x_i, y_j | x_{i-1}, x_{i-2}, y_{i-1}, y_{i-2}) \propto$$
$$P(x_i | x_{i-1}, x_{i-2}) P(y_j | y_{j-1}, y_{j-2}) P(x_i, y_j)$$

Transition probability in each language

"Coupling" probability: compatibility of tag pair

Essentially, a *product of experts*.

Bayesian Generative Story

Bayesian Generative Story

- For each language, draw:
  - Transition distributions over tag space (conditioned on previous two tags)
  - Emission distributions over lexicon (conditioned on tag)
- Draw coupling distribution over space of bilingual tag pairs

Bayesian Generative Story

- For each language, draw:
  - Transition distributions over tag space (conditioned on previous two tags)
  - Emission distributions over lexicon (conditioned on tag)
- Draw coupling distribution over space of bilingual tag pairs

All drawn from Dirichlet priors of appropriate dimension.

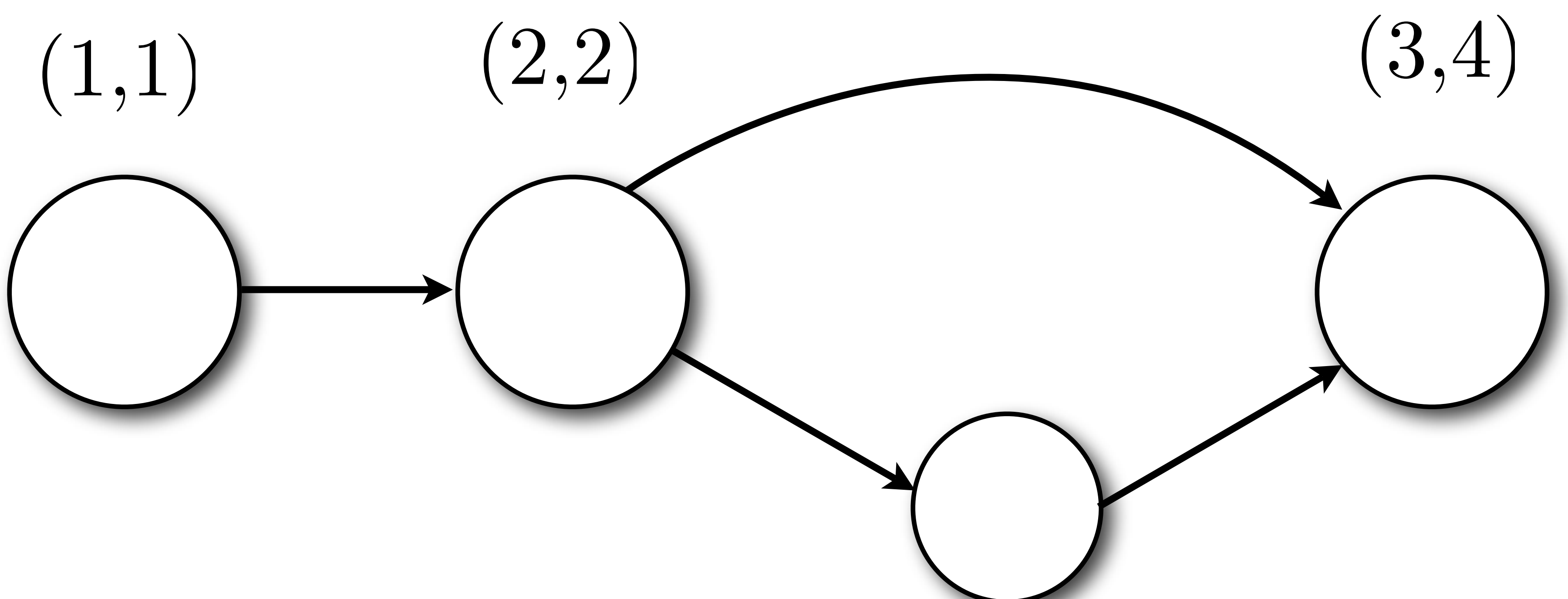
Bayesian Generative Story (cont'd)

For each bilingual parallel sentence:

Bayesian Generative Story
(cont'd)

For each bilingual parallel sentence:

1. Draw an *alignment*



Alignment must be 1-1 and contain no crossing edges

Treated as *observed variable* (based on GIZA++ alignments)

Bayesian Generative Story (cont'd)

<u>For each bilingual parallel sentence:</u>

1. Draw an *alignment*

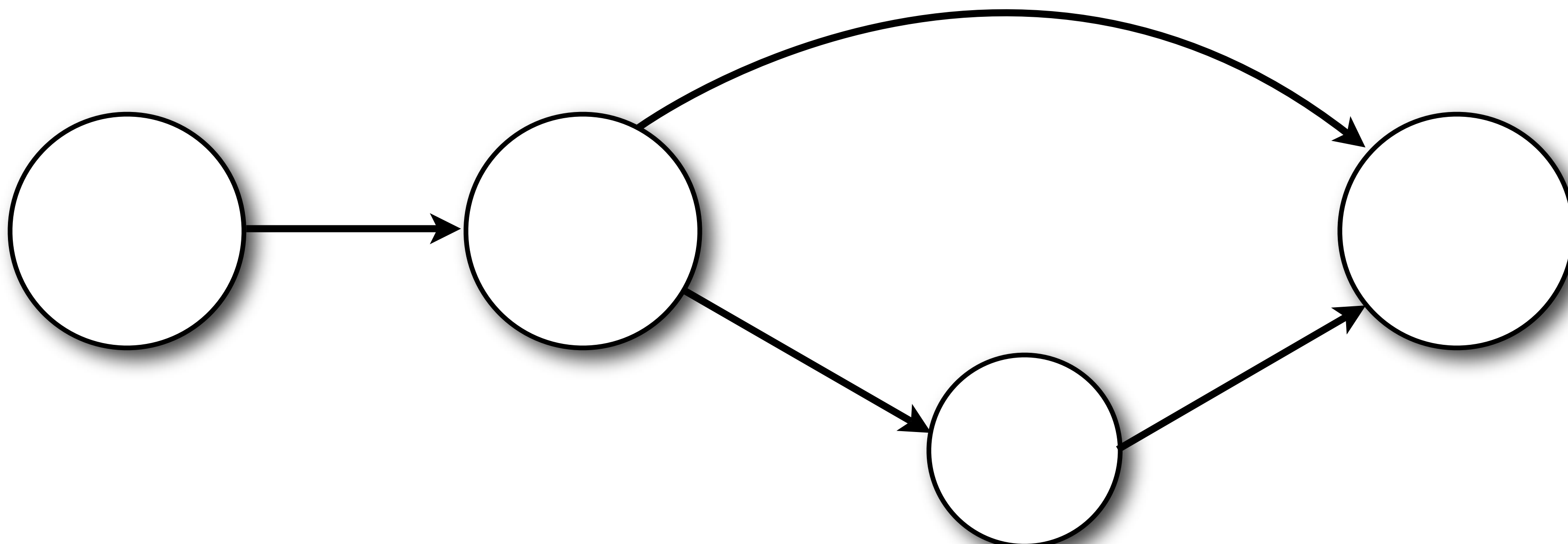
2. Draw parallel bilingual stream of tags in sequence from left to right

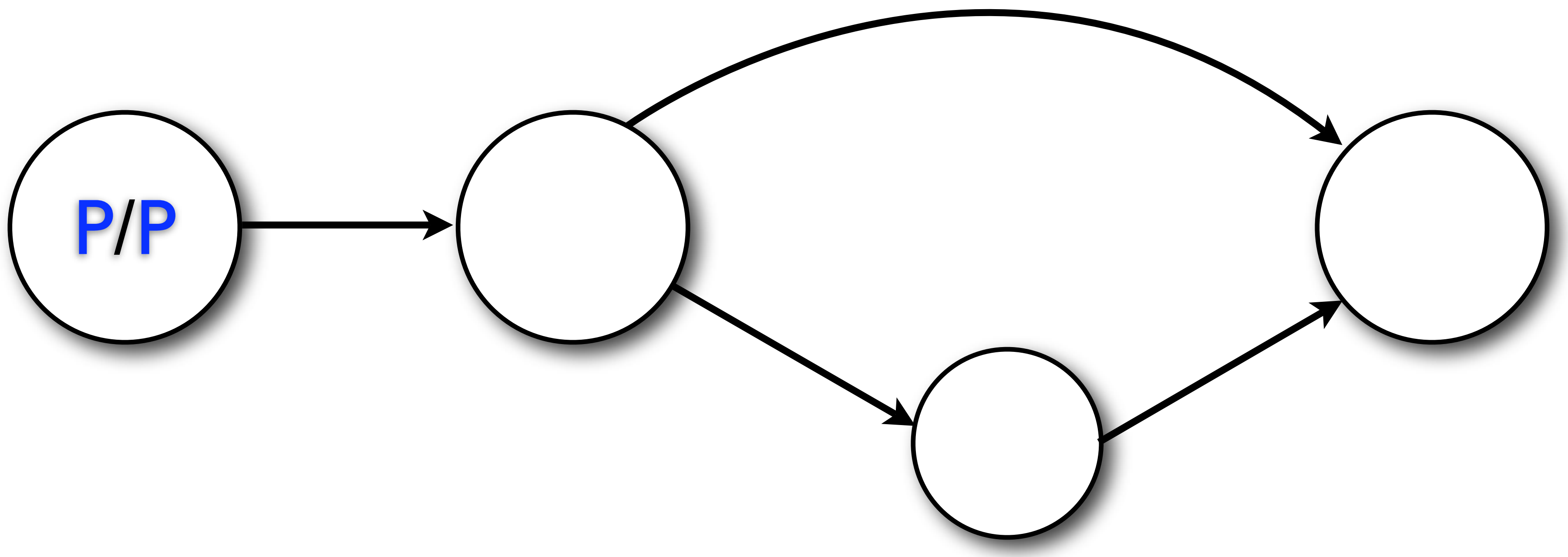
   - Unaligned tags drawn according to language-specific transition parameters

$$P(x_i|x_{i-1}, x_{i-2})$$

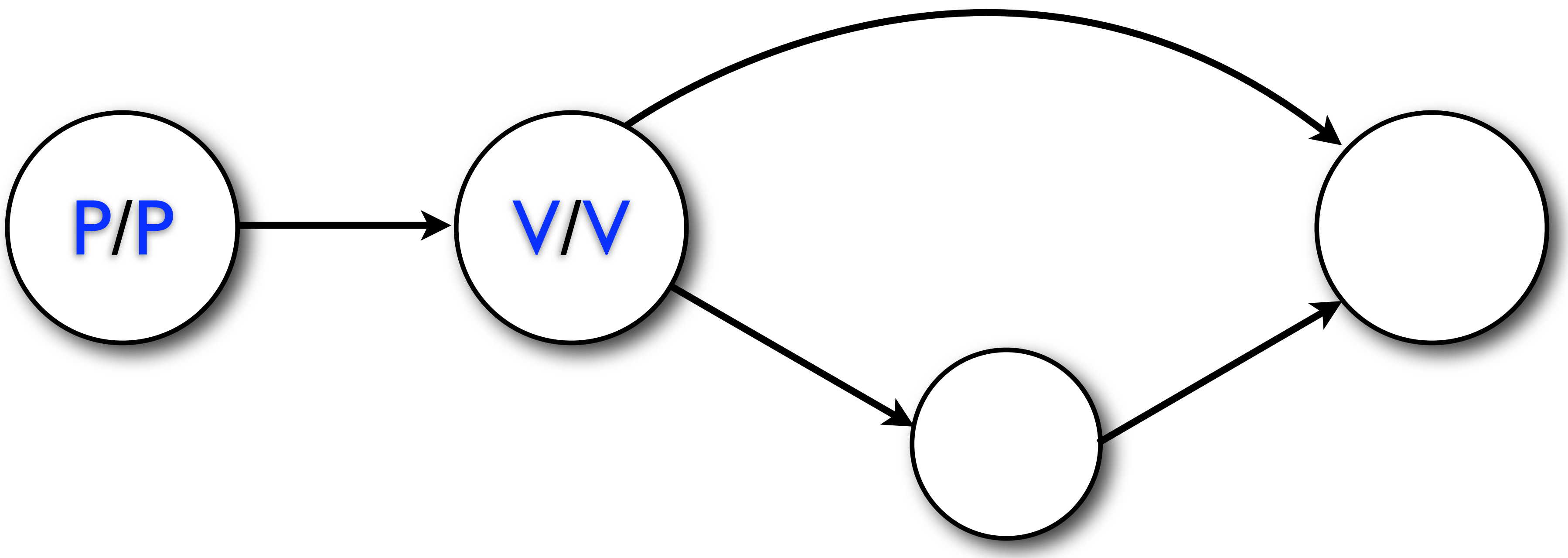
   - Aligned tag-pairs drawn jointly according to transitions and bilingual coupling parameter

$$\propto P(x_i|x_{i-1}, x_{i-2})P(y_j|y_{j-1}, y_{j-2})P(x_i, y_j)$$

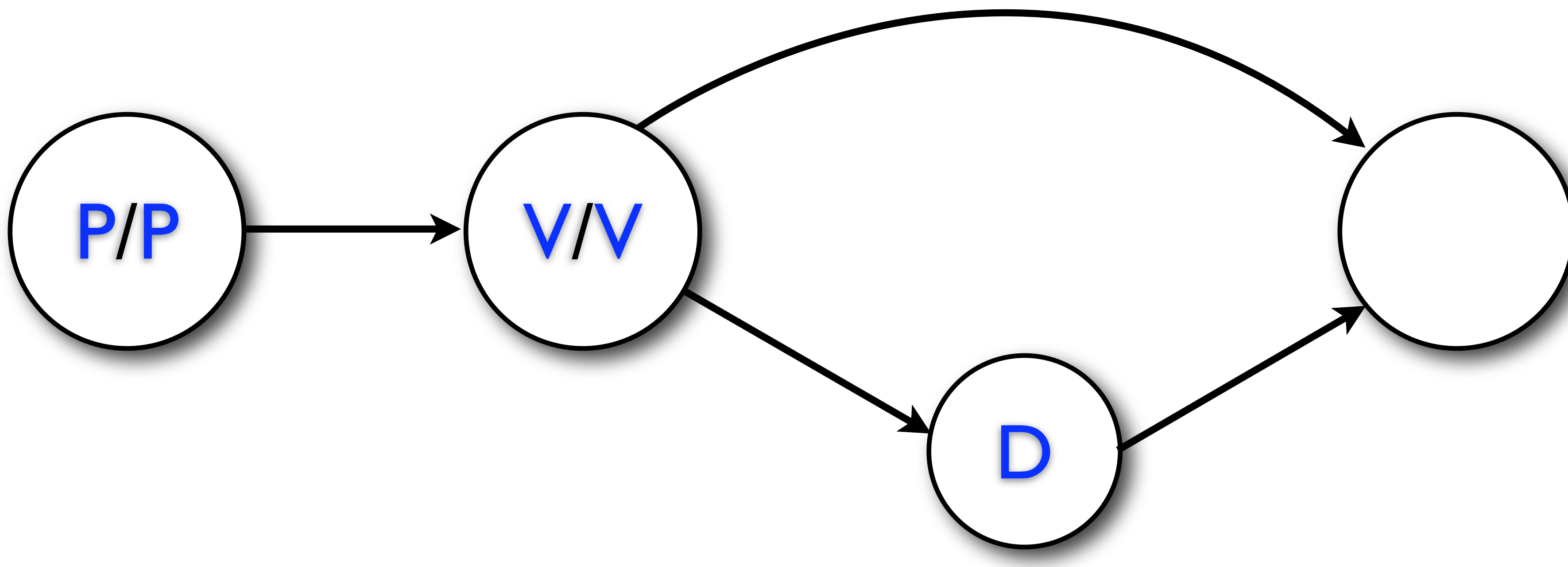




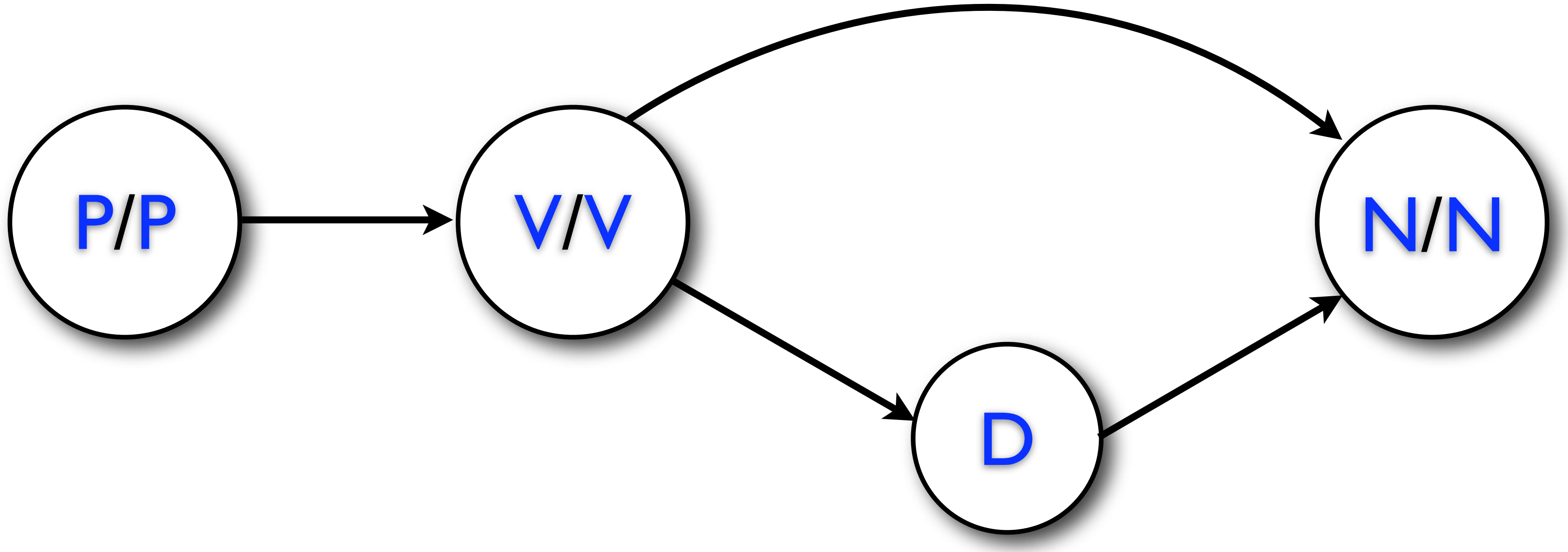
$$\propto trans_1(\mathrm{P}|\#\mathrm{START}) \cdot trans_2(\mathrm{P}|\#\mathrm{START}) \cdot coupling(\mathrm{P}, \mathrm{P})$$



$$\propto trans_1(\mathrm{V}|\mathrm{P}) \cdot trans_2(\mathrm{V}|\mathrm{V}) \cdot coupling(\mathrm{V}, \mathrm{V})$$



$$trans_2(\mathrm{D}|\mathrm{V})$$



$$\propto trans_1(\mathrm{N}|\mathrm{V}) \cdot trans_2(\mathrm{N}|\mathrm{D}) \cdot coupling(\mathrm{N}, \mathrm{N})$$

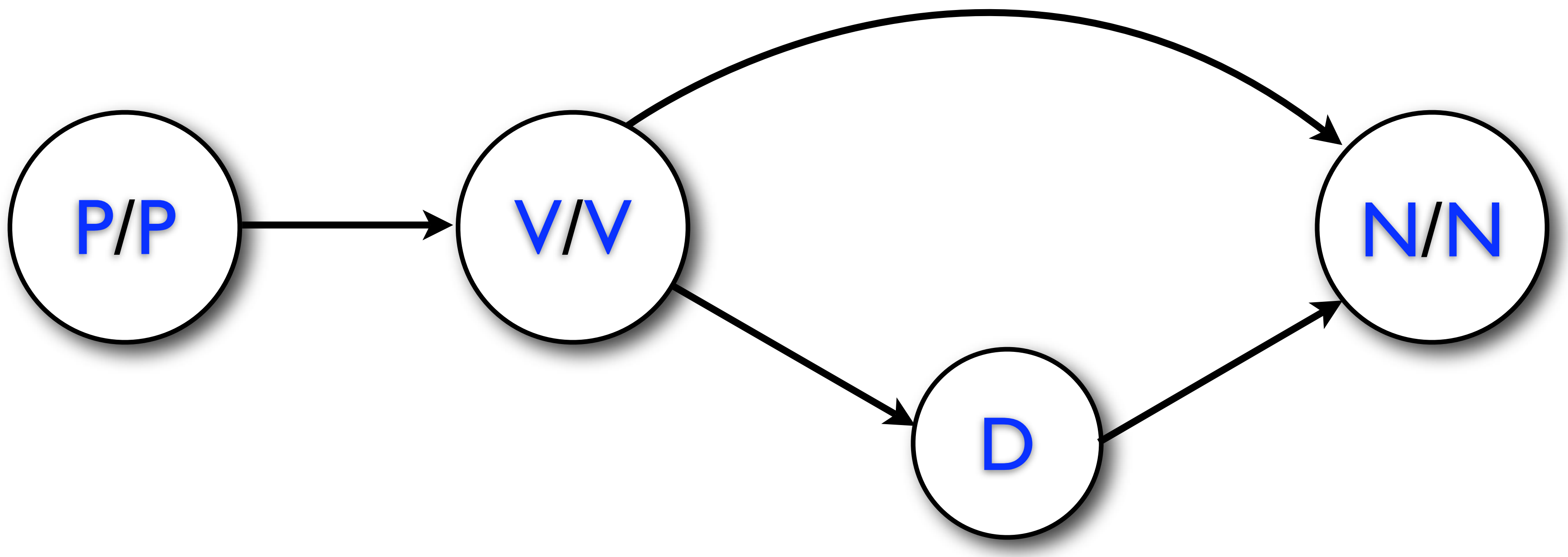
Bayesian Generative Story
(cont'd)

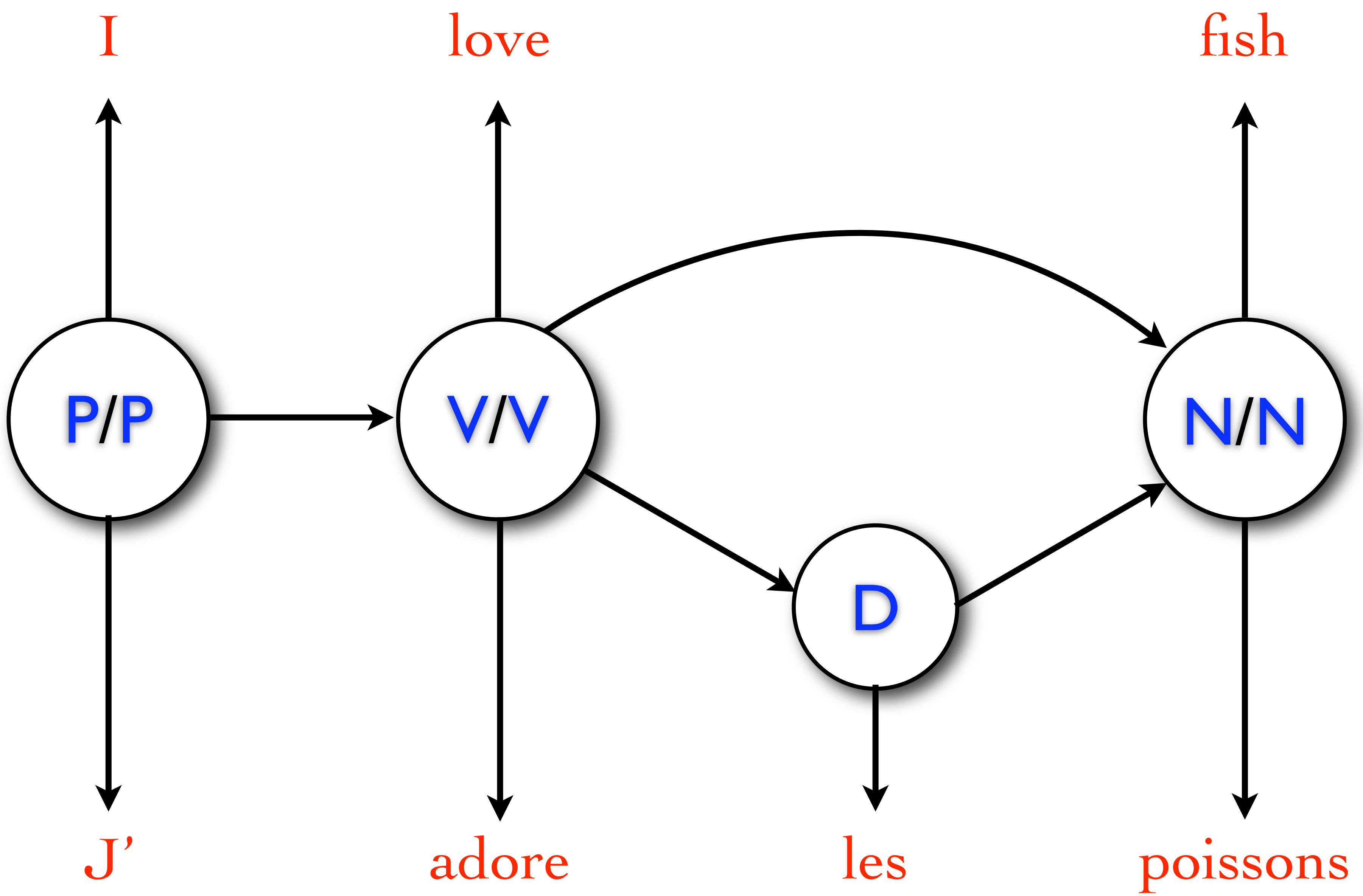
For each bilingual parallel sentence:

1. Draw an *alignment*

2. Draw parallel bilingual stream of tags in sequence from left to right

3. Draw words according to language-specific emission parameters.


P/P
V/V
N/N
D



$$emit_1(\text{“I”}|\text{P}) \cdot emit_2(\text{“J”}|\text{P}) \cdot \ldots$$

Bayesian Inference

Bayesian Inference

- Treat words and GIZA++ alignments as *observed variables*: x

Bayesian Inference

- Treat words and GIZA++ alignments as *observed variables*: x
- Treat emission, transition, and coupling parameters as *hidden variables*: θ

Bayesian Inference

- Treat words and GIZA++ alignments as *observed variables*: x
- Treat emission, transition, and coupling parameters as *hidden variables*: θ
- Predict POS tags y with highest posterior probability:

$$\operatorname*{argmax}_y P(y|x) = \operatorname*{argmax}_y \int_\theta P(y, x|\theta) P(\theta)\, d\theta$$

Sampling

Iteratively sample each variable conditioned on current value of others (Gibbs):

- Sample aligned tag-pairs and unaligned tags
- Sample* transition distributions
- Sample* coupling distribution

Sampling

Iteratively sample each variable conditioned on current value of others (Gibbs):

- Sample aligned tag-pairs and unaligned tags
- Sample* transition distributions
- Sample* coupling distribution

*no closed form using counts, due to factored parameterization:

$$P(x_i, y_j | \ldots) = \frac{P(x_i | x_{i-1}, x_{i-2}) P(y_j | y_{j-1}, y_{j-2}) P(x_i, y_j)}{Z}$$

Sampling

Iteratively sample each variable conditioned on current value of others (Gibbs):

- Sample aligned tag-pairs and unaligned tags
- Sample* transition distributions
- Sample* coupling distribution

*no closed form using counts, due to factored parameterization:

$$P(x_i, y_j | \ldots) = \frac{P(x_i | x_{i-1}, x_{i-2}) P(y_j | y_{j-1}, y_{j-2}) P(x_i, y_j)}{Z}$$

So we intersperse Gibbs with a Metropolis-Hastings step

Metropolis-Hastings

- Define tractable proposal distribution: Q
- Sample a new value: $z^* \sim Q$
- Accept with probability: $min\left\{1, \frac{P(z^*)Q(z)}{P(z)Q(z^*)}\right\}$

Metropolis-Hastings

- Define tractable proposal distribution: Q
- Sample a new value: $z^* \sim Q$
- Accept with probability: $min\left\{1, \frac{P(z^*)Q(z)}{P(z)Q(z^*)}\right\}$

For the coupling distribution, we use proposal:

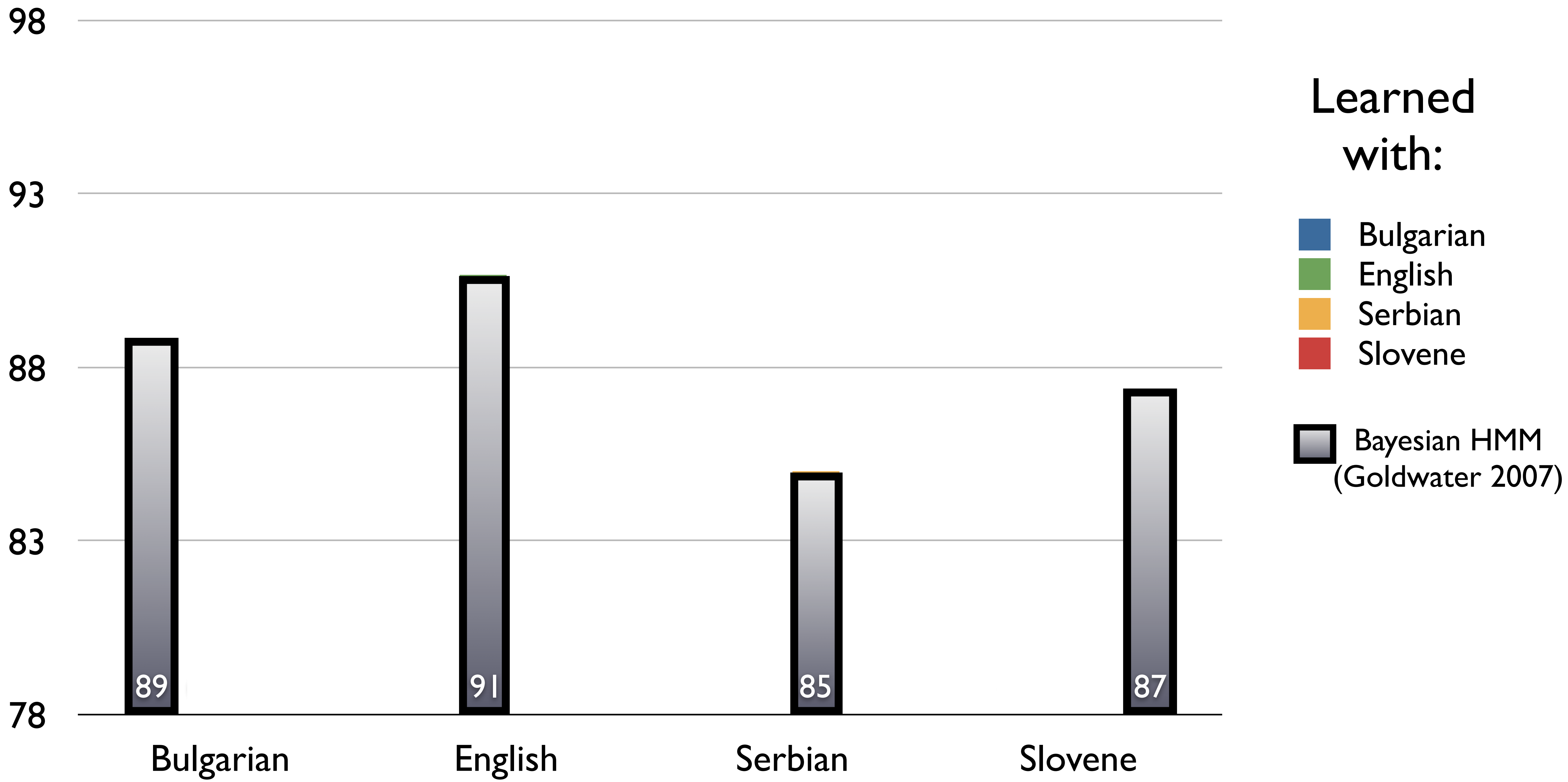
$$Q \equiv \mathrm{Dir}(Count(N,N), Count(N,V), ...)$$

Counts of coupled parts-of-speech
according to current sampled tags

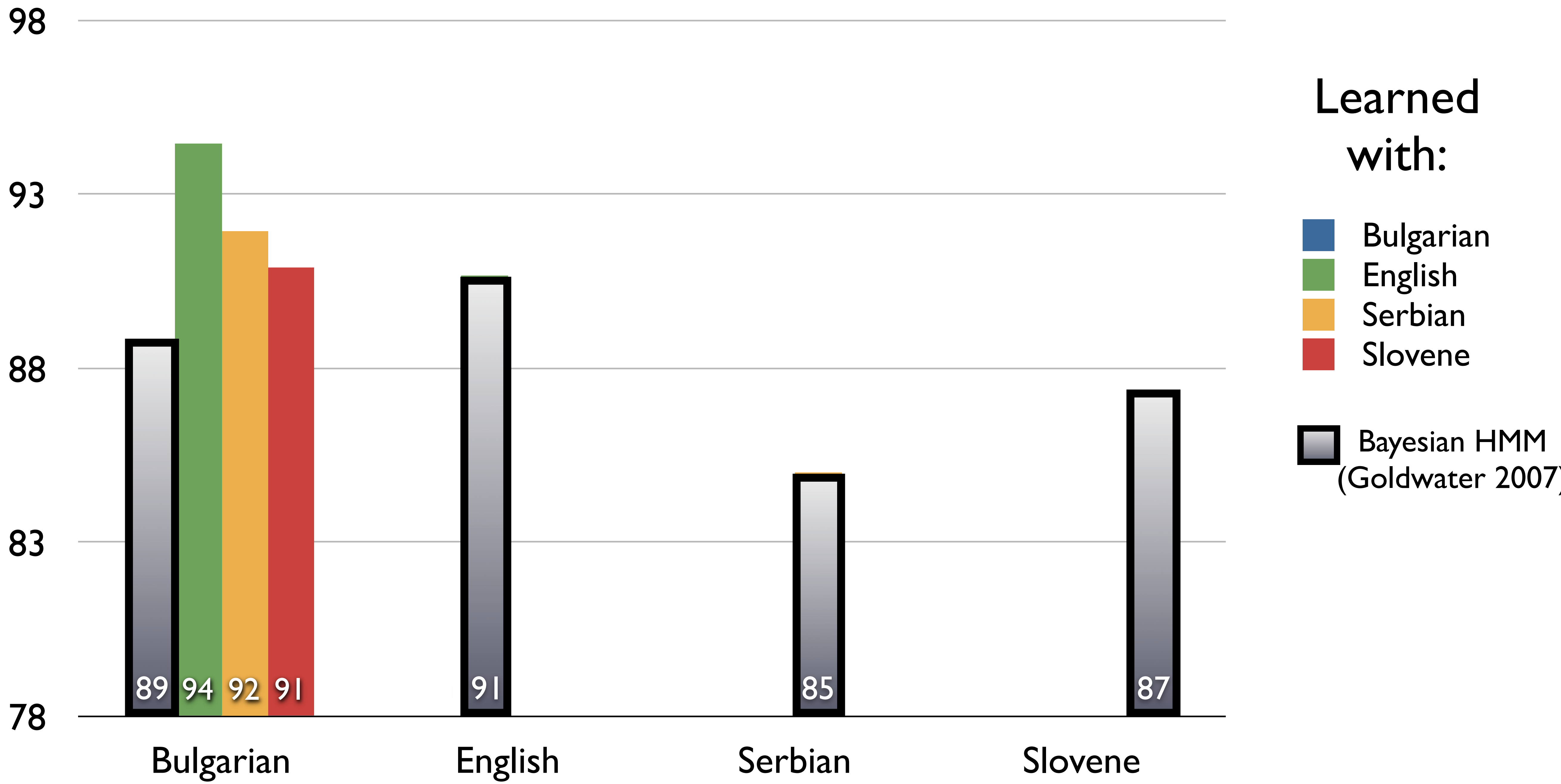
Evaluation Setup

- Evaluate on *monolingual* test-set
- Orwell's Nineteen Eighty Four
  - Languages: English, Bulgarian, Serbian, Slovene
  - 94,725 tokens (English)
  - 13 coarse POS tags (Multext East corpus)
- GIZA++ alignments
  - Intersection of each direction (1-1)
  - Removal of crossing edges (< 5%)

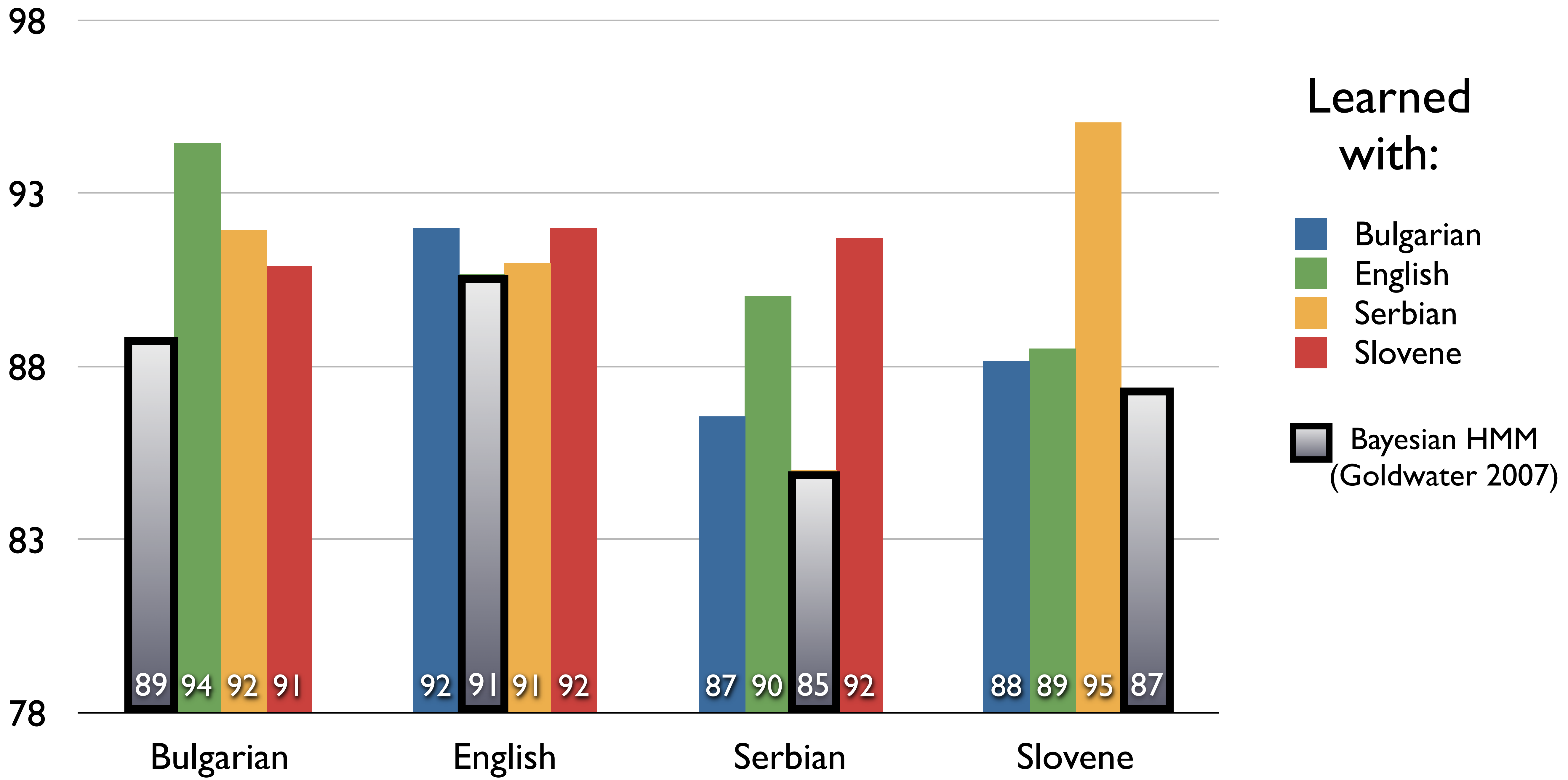
Accuracy
(full lexicon)



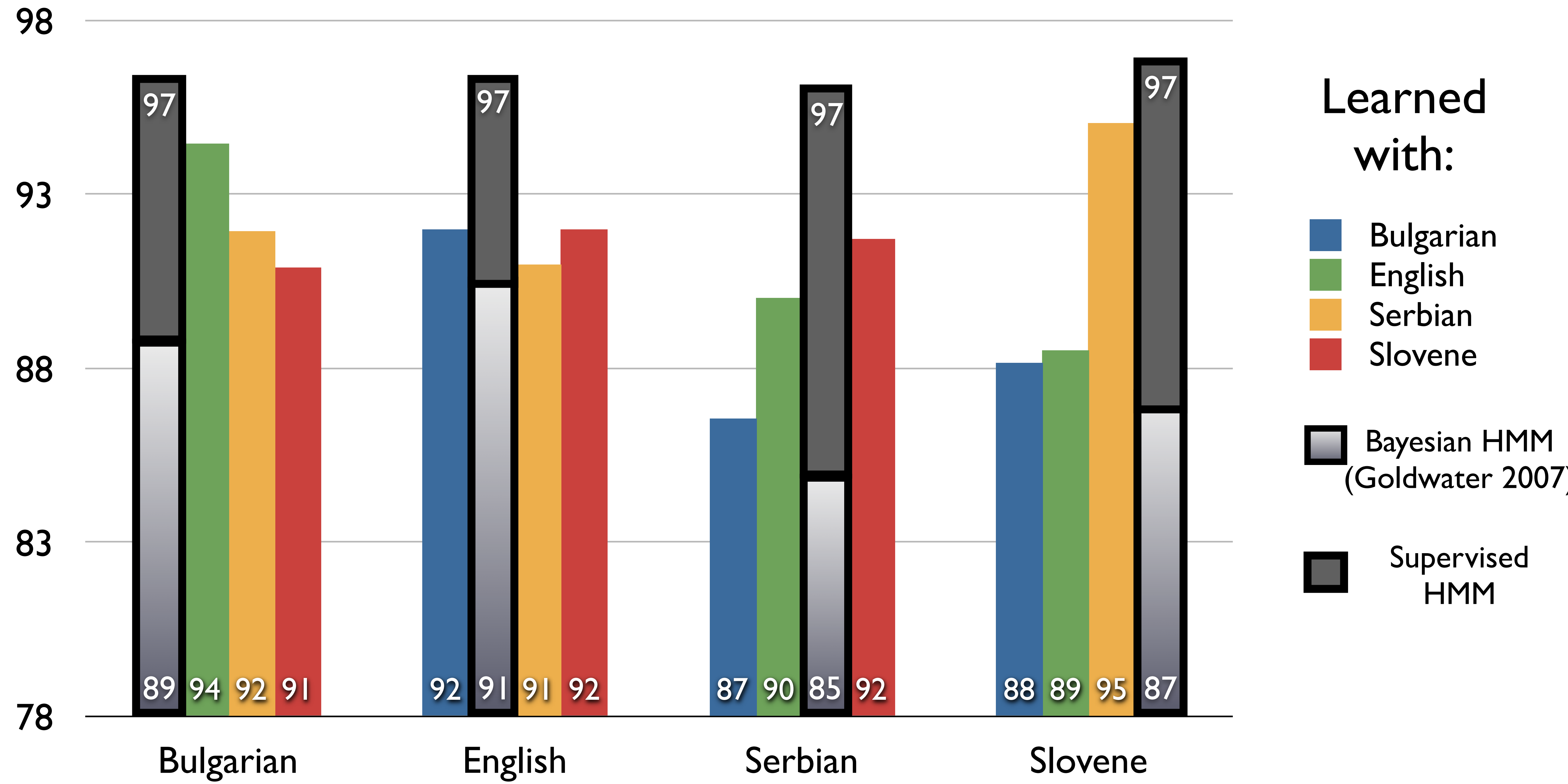
Accuracy
(full lexicon)



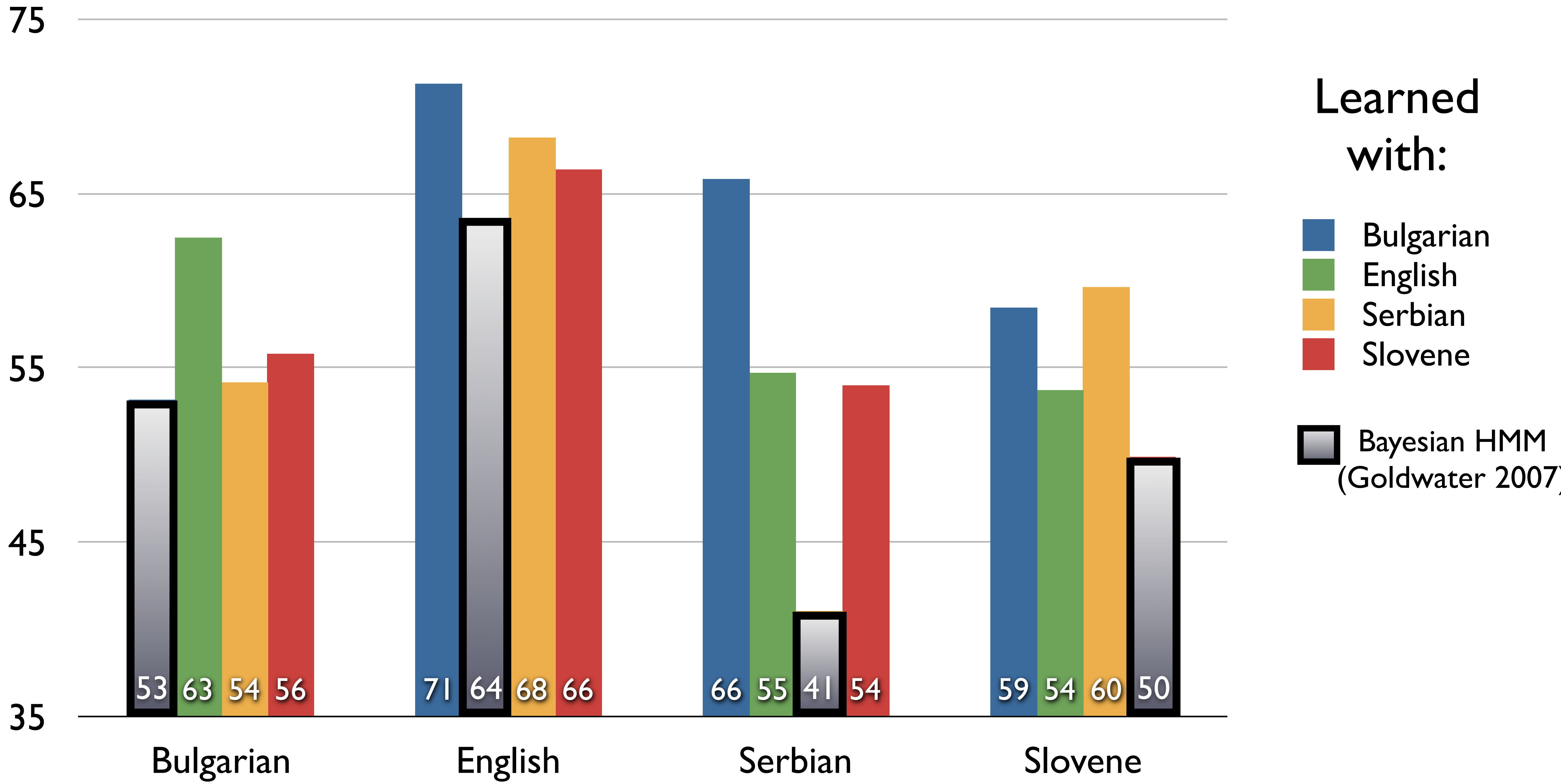
Accuracy
(full lexicon)



Accuracy
(full lexicon)



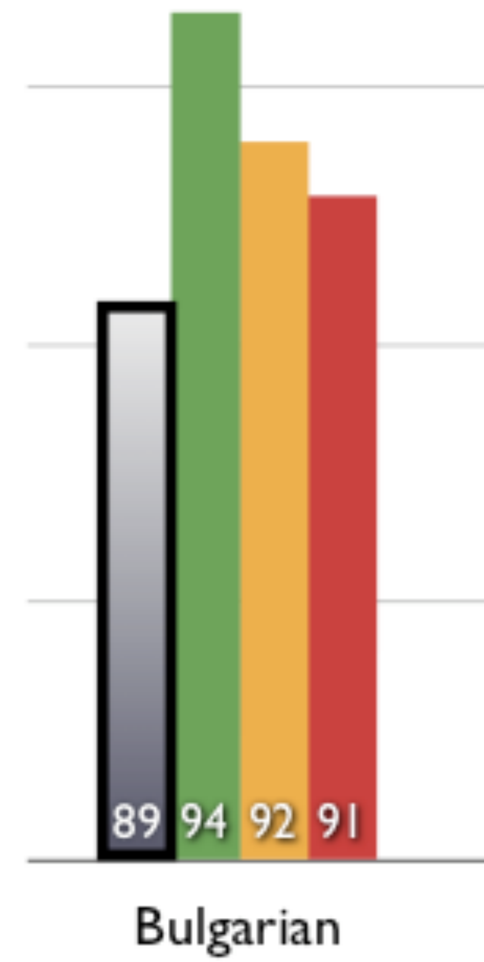
Accuracy
(100 word lexicon)



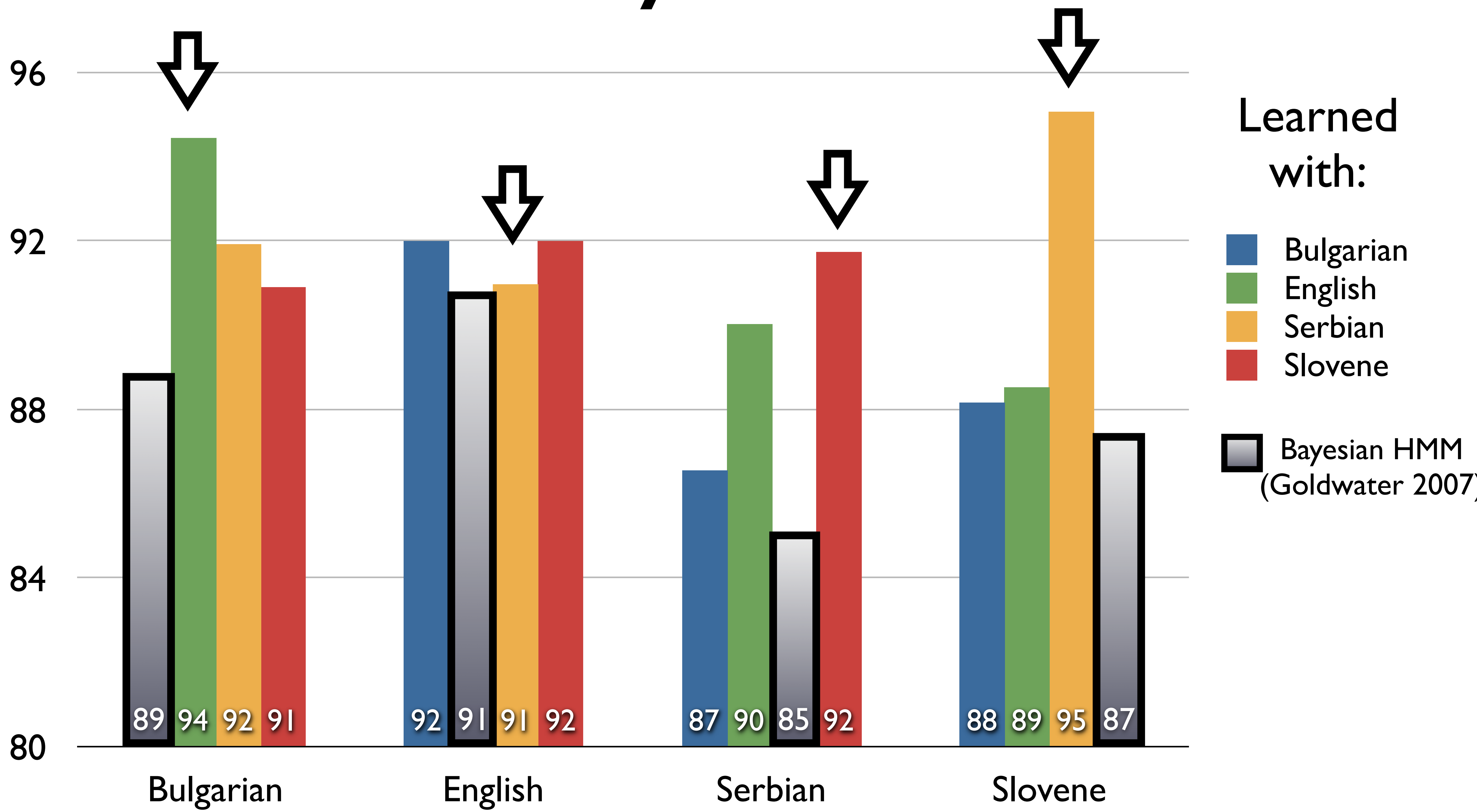
Cross-lingual Analysis

- Some language pairings much better than others (Serbian + Slovene, English + Bulgarian)
- Given gold tags, easy to predict relative performance gains using cross-lingual entropy:

$$H\left[P(x_i | y_j, (i,j) \in a)\right]$$



Accuracy (full lexicon)



lowest cross-lingual entropy

Open Question

Open Question

How to predict optimal pairings in *unsupervised* manner?

Open Question

How to predict optimal pairings in *unsupervised* manner?

- Family relatedness not accurate predictor

Open Question

How to predict optimal pairings in *unsupervised* manner?

- Family relatedness not accurate predictor
- Typological relatedness..?

Open Question

How to predict optimal pairings in *unsupervised* manner?

- Family relatedness not accurate predictor
- Typological relatedness..?
  - English & Bulgarian analytical, fixed word order

Open Question

How to predict optimal pairings in *unsupervised* manner?

- Family relatedness not accurate predictor
- Typological relatedness..?
  - English & Bulgarian analytical, fixed word order
  - Serbian & Slovene inflectional, variable word order

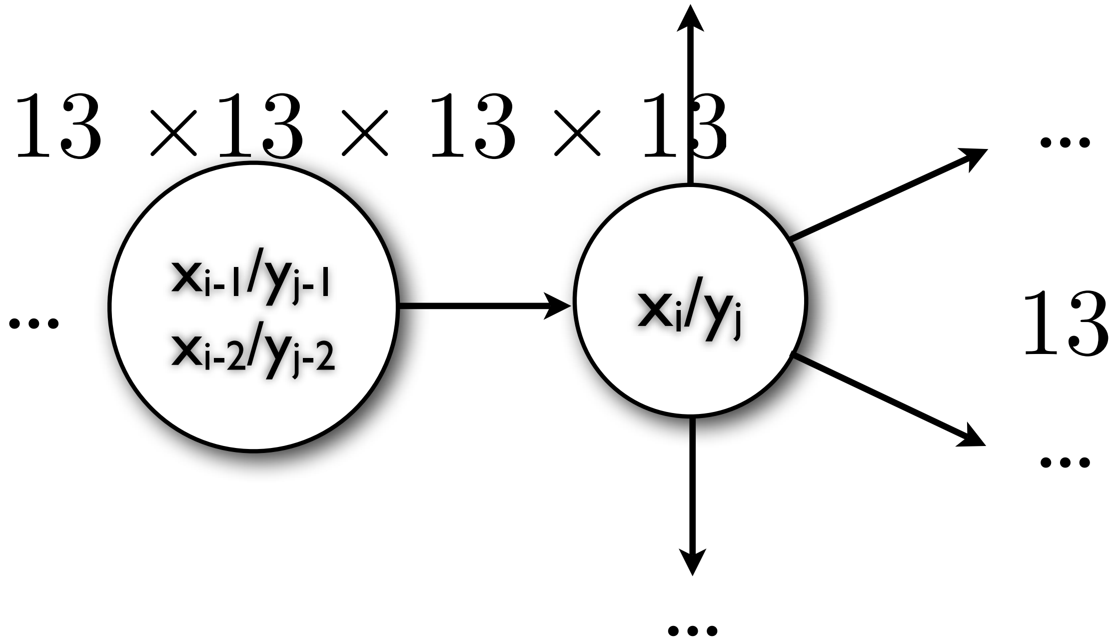
Conclusions

- Unsupervised multilingual learning effective for POS tagging.
- Beneficial for *all* pairings, drastic improvement for some.
- <u>Unsupervised/Supervised gap</u>:
    - Avg over all pairings: cut by 1/3.
    - Using best pairings: cut by 1/2.

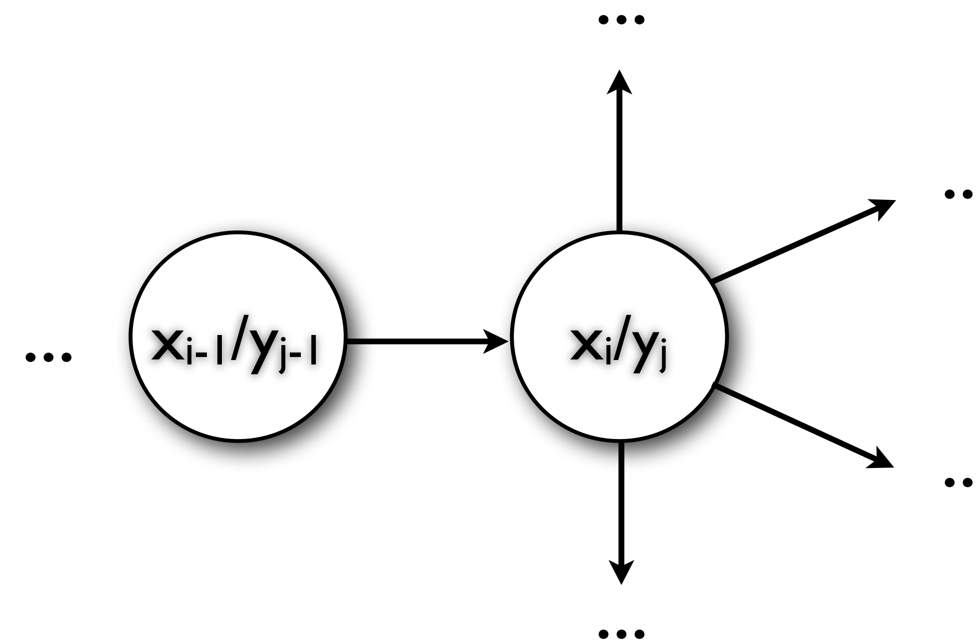
88	91	93	97

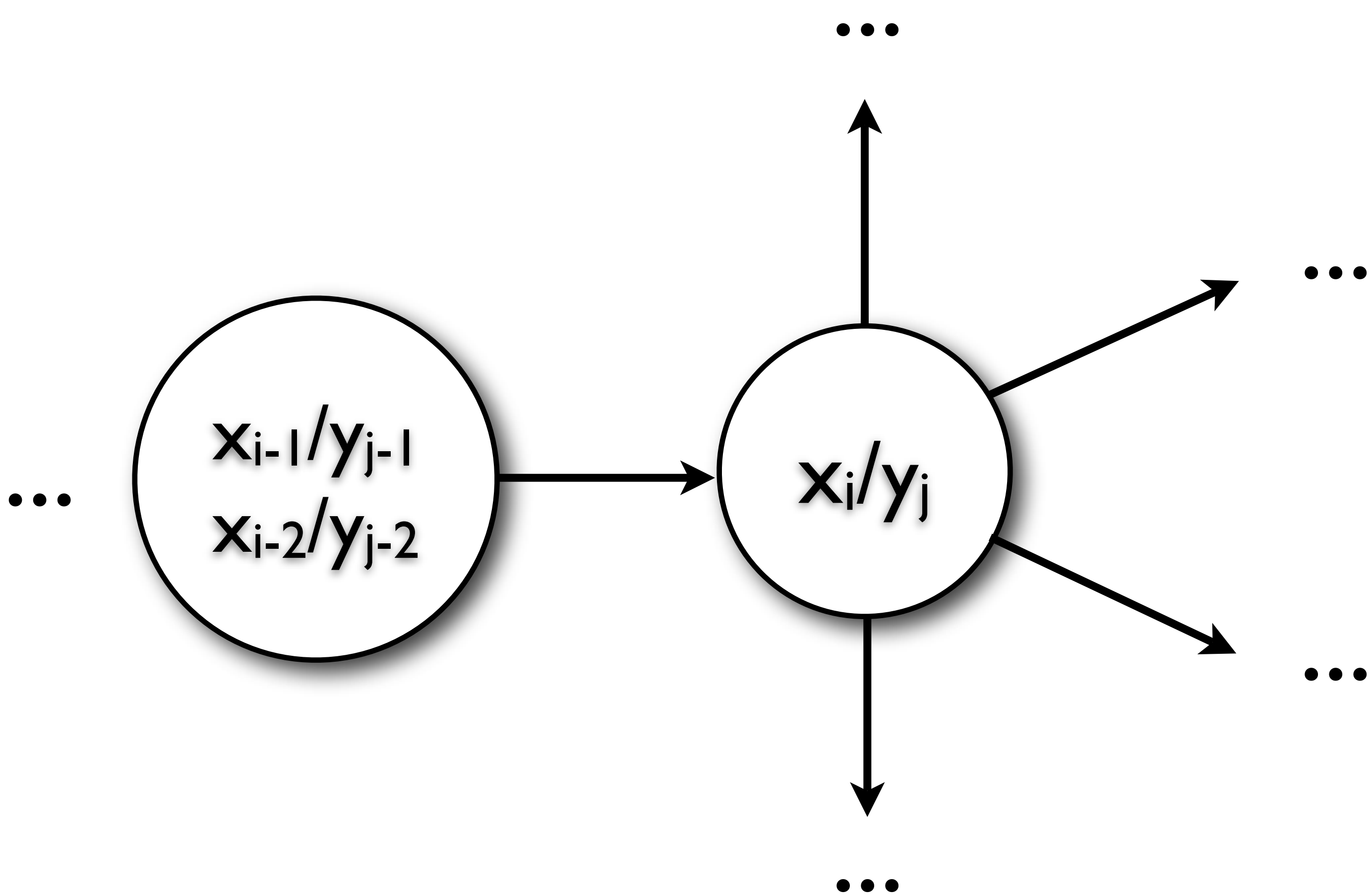
(full lexicon experiment)

$13 \times 13 \times 13 \times 13$



13×13

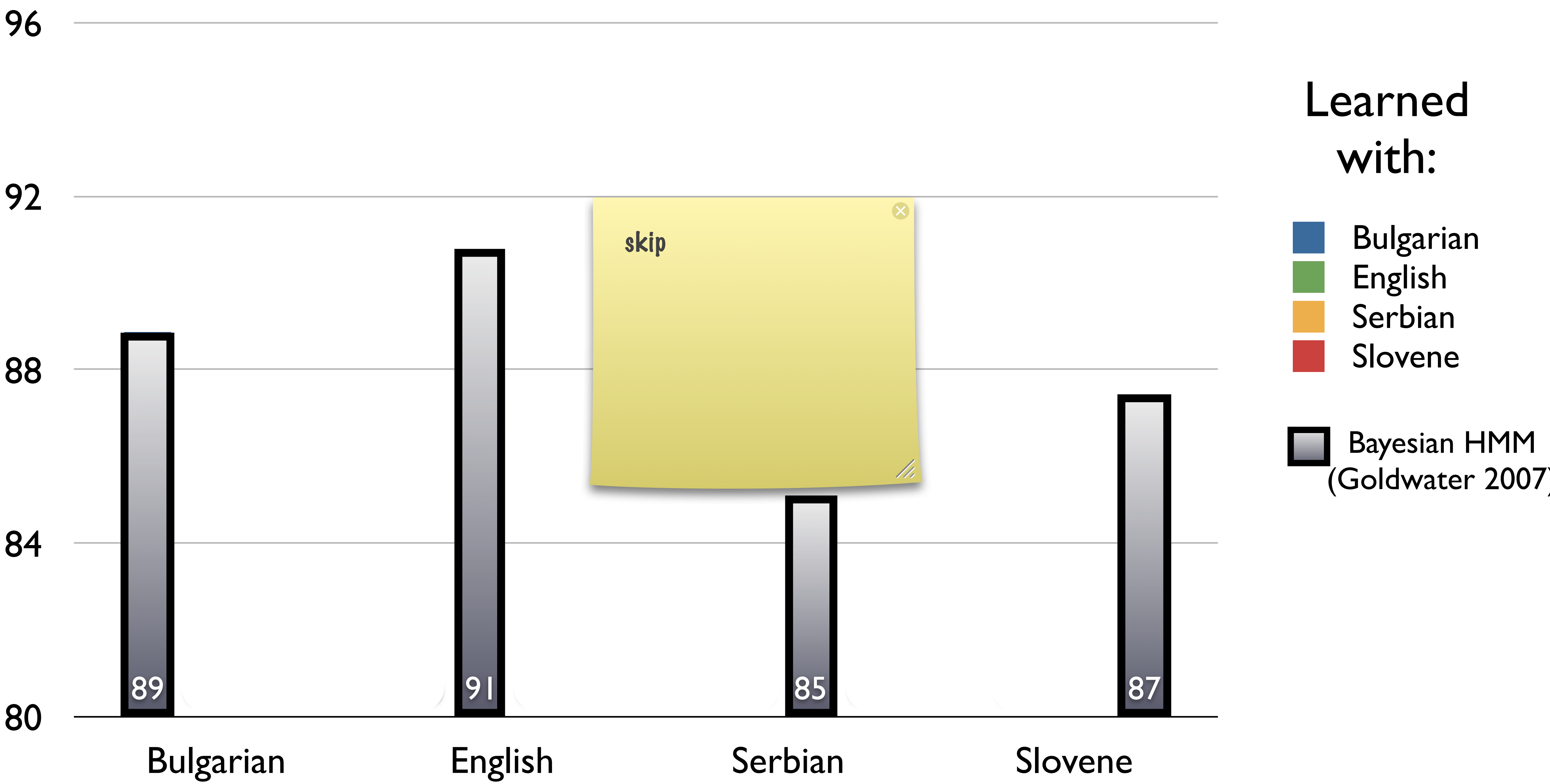




[P]	[V]		[N]
I	love		fish.
J’	aime	les	poissons.
[P]	[V]	[D]	[N]

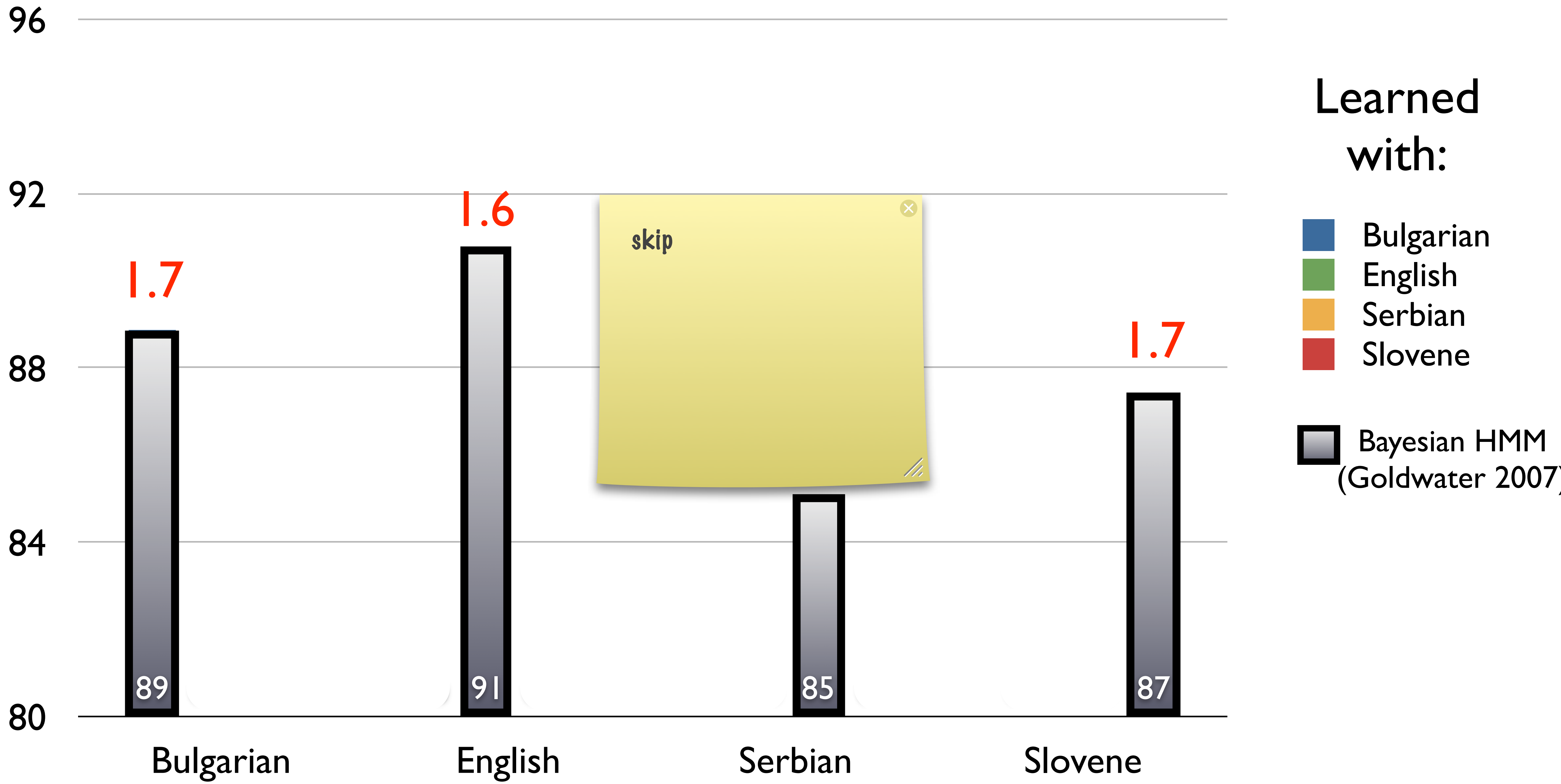
			[V]
I	love	to	fish.
J’	aime		pêcher.
			[V]

Accuracy
(full lexicon)



average trigram entropy: $H\left[P(x_i | x_{i-1}, x_{i-2})\right]$

Accuracy
(full lexicon)

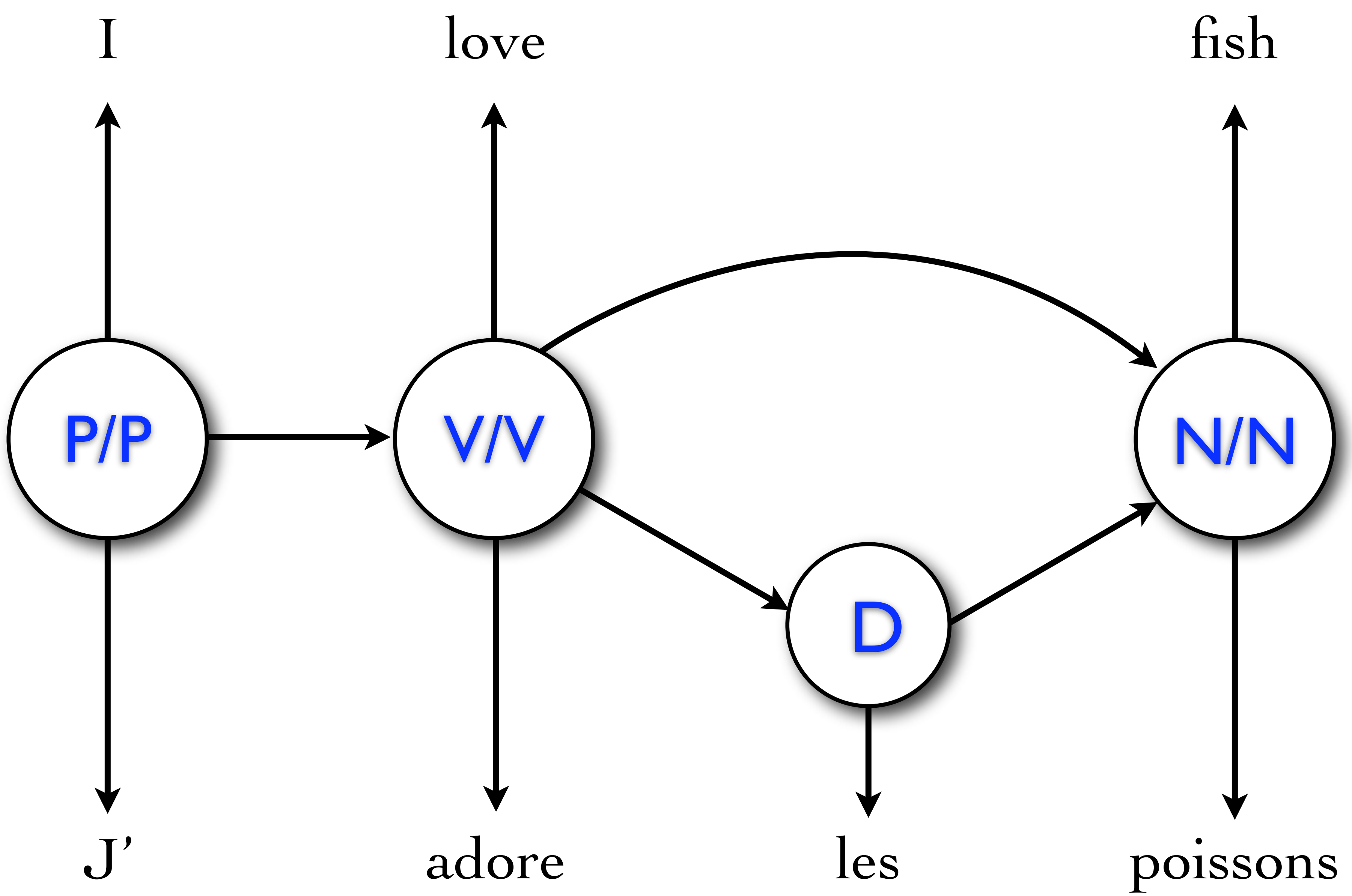


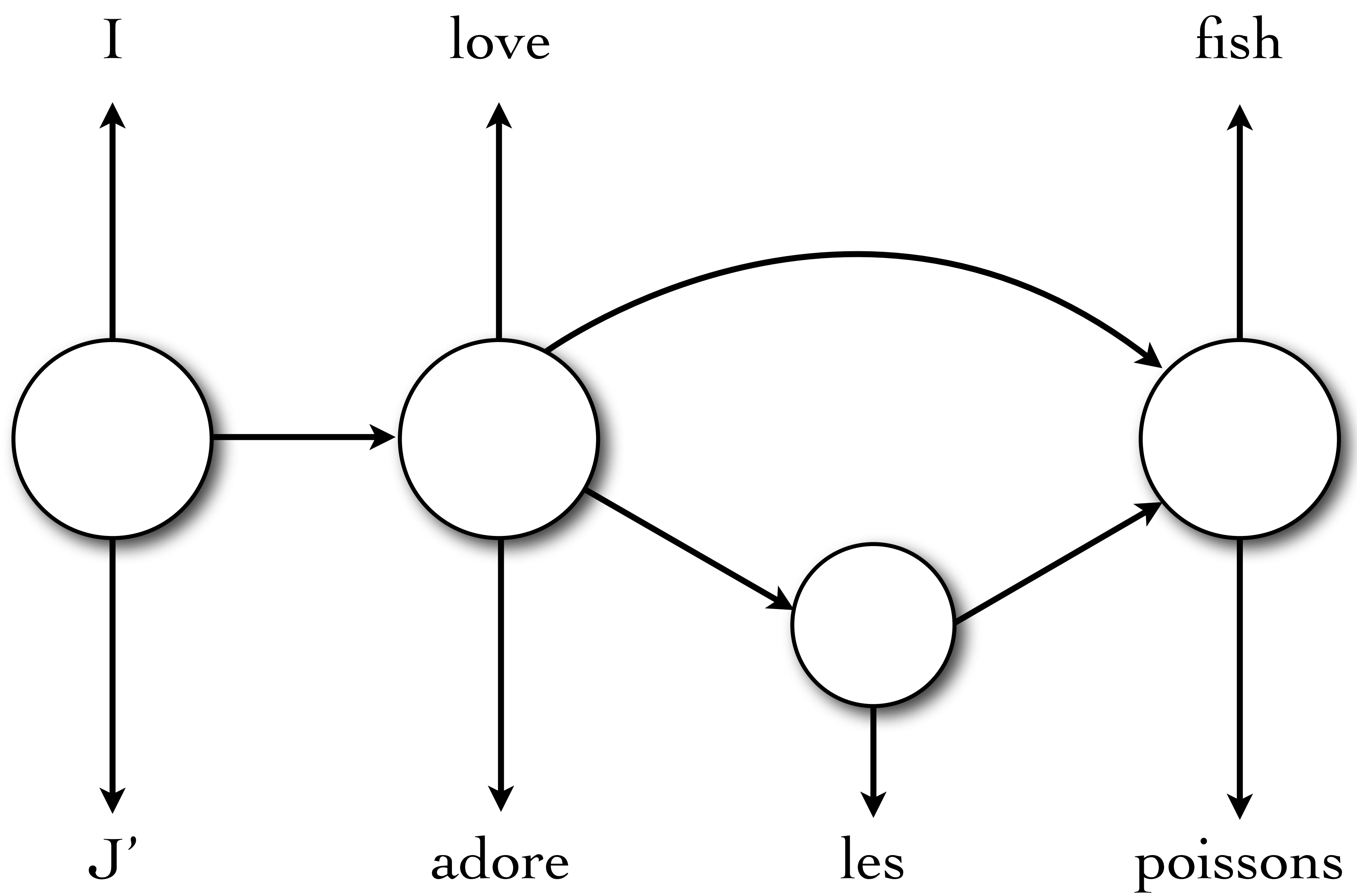
average trigram entropy: $H\left[P(x_i | x_{i-1}, x_{i-2})\right]$

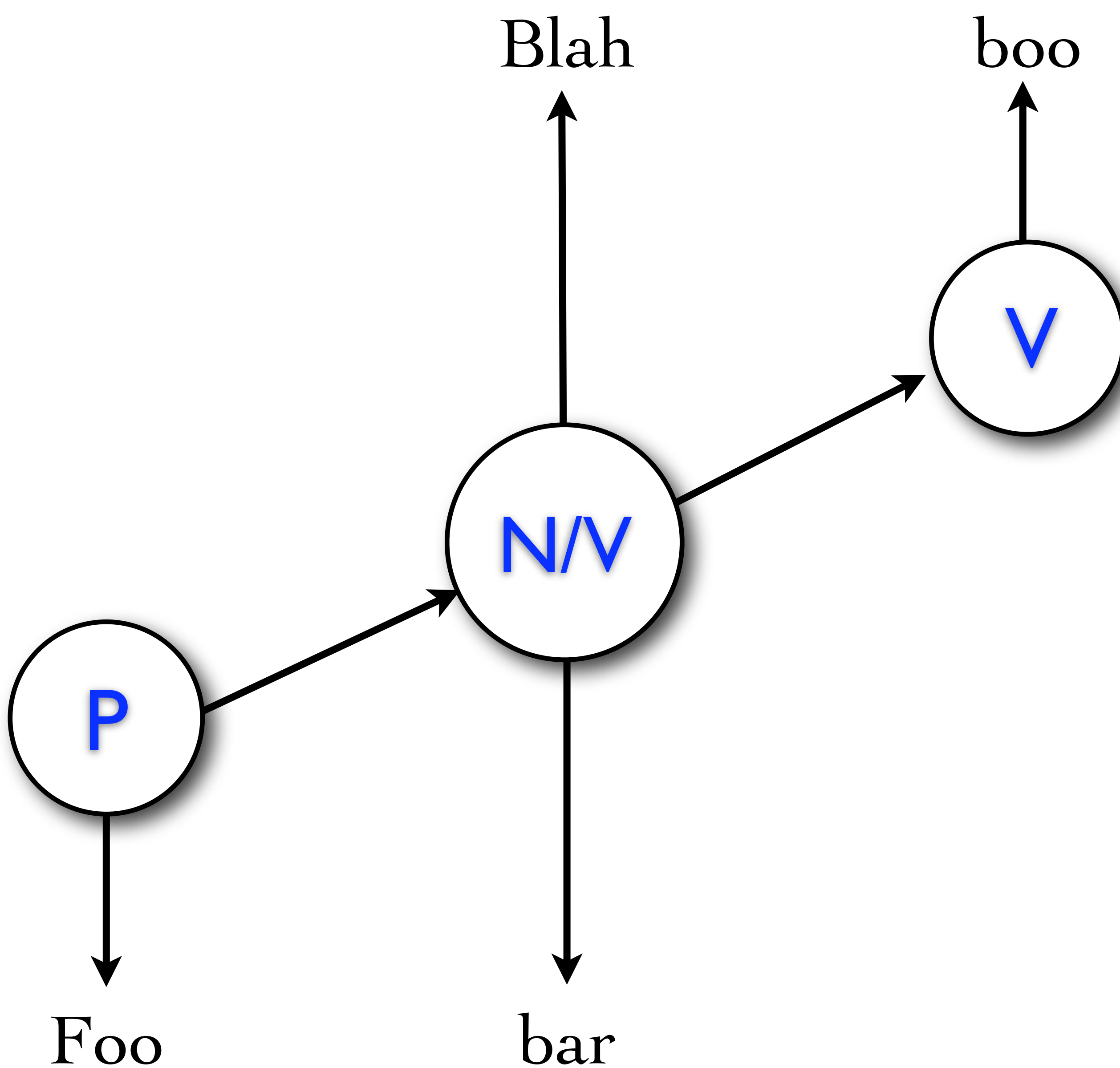
Tagset

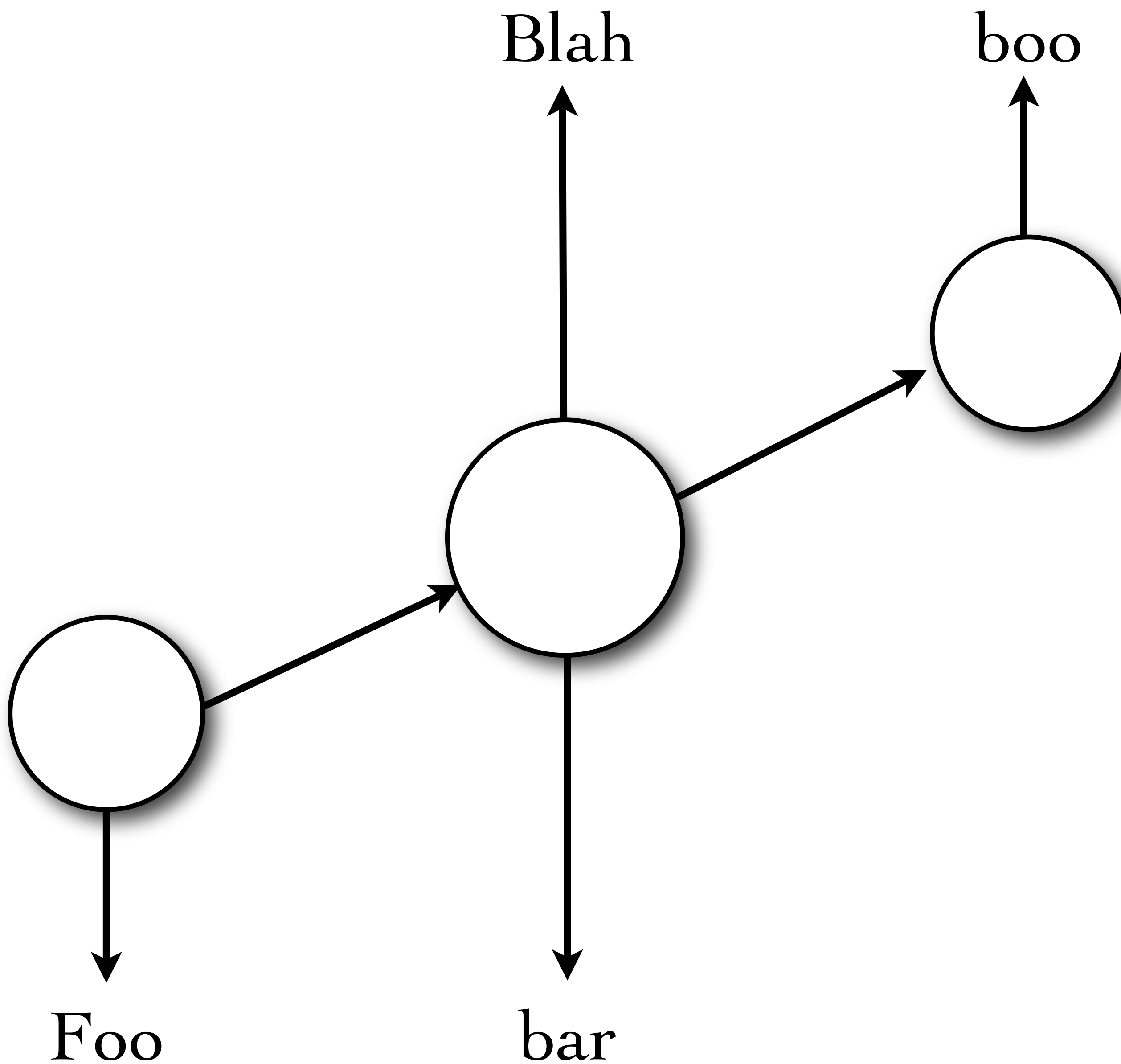
- Gold Standard: Multext-East Corpus
- Tag repository: 13 categories
- Tags/Token Ratio in corpus

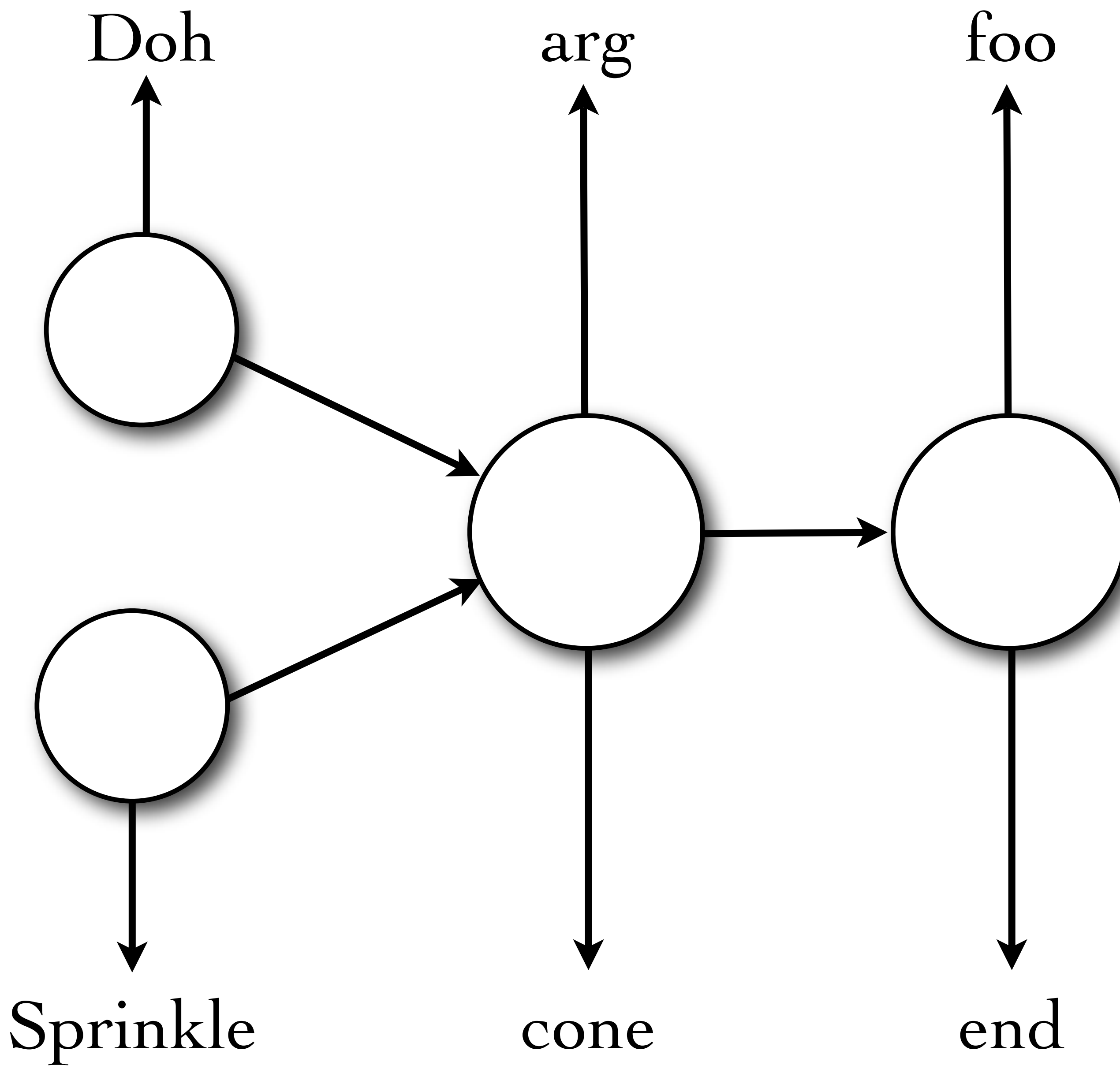
Language	Tag/Token
Serbian	1.41
Slovene	1.40
Bulgarian	1.34
English	2.58

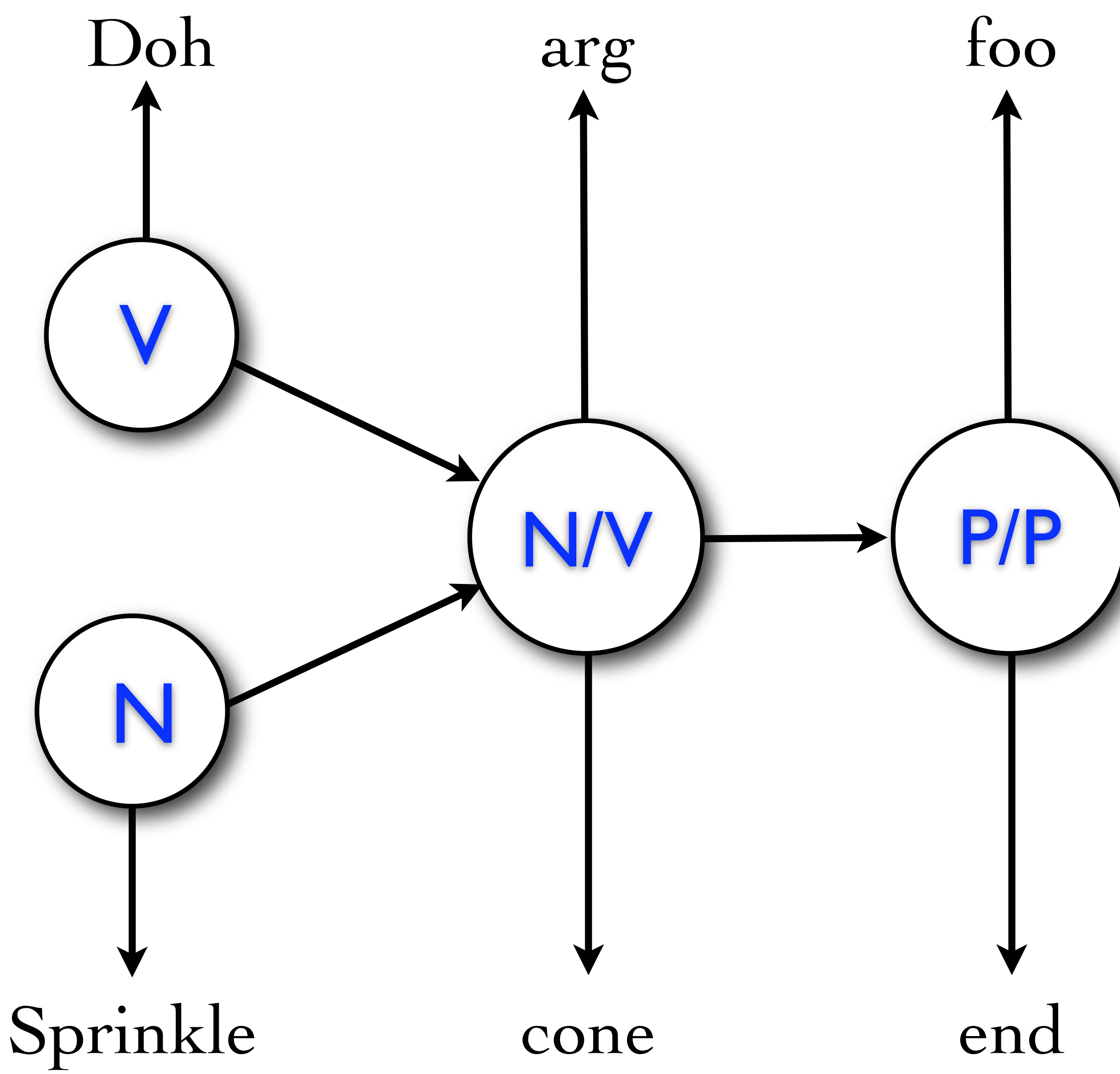

I
love
fish
P/P
V/V
N/N
D
J’
adore
les
poissons


I
love
fish
J'
adore
les
poissons


Blah
boo
V
N/V
P
Foo
bar


Blah
boo
Foo
bar


Doh
arg
foo
Sprinkle
cone
end


Doh
arg
foo
V
N/V
P/P
N
Sprinkle
cone
end