Model Combination for Machine Translation

Google research

John DeNero, Shankar Kumar,
Ciprian Chelba, and Franz Josef Och

Motivation

- A statistical machine translation model scores derivations

  (log) model score sums language model and translation model

Motivation

- A statistical machine translation model scores derivations

(log) model score sums language model and translation model

$$\theta \cdot \phi(d) = \theta \cdot \left[\sum_{w \in \text{n-grams}(d)} \phi_{\text{LM}}(w) + \sum_{r \in \text{rules}(d)} \phi_{\text{TM}}(r) \right]$$

Motivation

- A statistical machine translation model scores derivations

  (log) model score sums language model and translation model

$$\theta \cdot \phi(d) = \theta \cdot \left[\sum_{w \in \text{n-grams}(d)} \phi_{\text{LM}}(w) + \sum_{r \in \text{rules}(d)} \phi_{\text{TM}}(r) \right]$$

Motivation

- A statistical machine translation model scores derivations

(log) model score sums language model and translation model

$$\theta \cdot \phi(d) = \theta \cdot \left[\sum_{w \in \text{n-grams}(d)} \phi_{\text{LM}}(w) + \sum_{r \in \text{rules}(d)} \phi_{\text{TM}}(r) \right]$$

Motivation

- A statistical machine translation model scores derivations

  (log) model score sums language model and translation model

$$\theta \cdot \phi(d) = \theta \cdot \left[\underbrace{\sum_{w \in \text{n-grams}(d)} \phi_{\text{LM}}(w)} + \underbrace{\sum_{r \in \text{rules}(d)} \phi_{\text{TM}}(r)} \right]$$

- We can improve over max-derivation decoding

Motivation

- A statistical machine translation model scores derivations

(log) model score sums language model and translation model

$$\theta \cdot \phi(d) = \theta \cdot \left[\sum_{w \in \text{n-grams}(d)} \phi_{\text{LM}}(w) + \sum_{r \in \text{rules}(d)} \phi_{\text{TM}}(r) \right]$$

- We can improve over max-derivation decoding
  - Consensus decoding (e.g., minimum Bayes risk)

Motivation

- A statistical machine translation model scores derivations

(log) model score sums language model and translation model

$$\theta \cdot \phi(d) = \theta \cdot \left[\sum_{w \in \text{n-grams}(d)} \phi_{\text{LM}}(w) + \sum_{r \in \text{rules}(d)} \phi_{\text{TM}}(r) \right]$$

- We can improve over max-derivation decoding
  - Consensus decoding (e.g., minimum Bayes risk) $\sum$

Motivation

- A statistical machine translation model scores derivations

(log) model score sums language model and translation model

$$\theta \cdot \phi(d) = \theta \cdot \left[\sum_{w \in \text{n-grams}(d)} \phi_{\text{LM}}(w) + \sum_{r \in \text{rules}(d)} \phi_{\text{TM}}(r) \right]$$

- We can improve over max-derivation decoding
  - Consensus decoding (e.g., minimum Bayes risk) $\sum$
  - System combination (e.g., confusion networks)

Motivation

- A statistical machine translation model scores derivations

  (log) model score sums language model and translation model

$$\theta \cdot \phi(d) = \theta \cdot \left[\sum_{w \in \text{n-grams}(d)} \phi_{\text{LM}}(w) + \sum_{r \in \text{rules}(d)} \phi_{\text{TM}}(r) \right]$$

- We can improve over max-derivation decoding
  - Consensus decoding (e.g., minimum Bayes risk) $\sum$
  - System combination (e.g., confusion networks)

Motivation

- A statistical machine translation model scores derivations

  (log) model score sums language model and translation model

$$\theta \cdot \phi(d) = \theta \cdot \left[\sum_{w \in \text{n-grams}(d)} \phi_{\text{LM}}(w) + \sum_{r \in \text{rules}(d)} \phi_{\text{TM}}(r) \right]$$

- We can improve over max-derivation decoding
  - Consensus decoding (e.g., minimum Bayes risk)
  - System combination (e.g., confusion networks)

- In this work, we develop a technique that integrates both

Consensus Decoding

- Derivation scores can be interpreted as probabilities

$$\mathrm{P}(d|f) = \frac{\exp\left(\theta \cdot \phi(d)\right)}{\sum_{d'} \exp\left(\theta \cdot \phi(d')\right)}$$

Consensus Decoding

- Derivation scores can be interpreted as probabilities

$$\mathrm{P}(d|f) = \frac{\exp\left(\theta \cdot \phi(d)\right)}{\sum_{d'} \exp\left(\theta \cdot \phi(d')\right)}$$

- We can query this distribution for:

Consensus Decoding

- Derivation scores can be interpreted as probabilities

$$\mathrm{P}(d|f) = \frac{\exp\left(\theta \cdot \phi(d)\right)}{\sum_{d'} \exp\left(\theta \cdot \phi(d')\right)}$$

- We can query this distribution for:

Whole translations [Blunsom et al., '08]:

$$e_{\text{MAX-TRANS}} = \arg\max_{e} \sum_{d:\sigma_e(d)=e} \mathrm{P}(d|f)$$

Consensus Decoding

- Derivation scores can be interpreted as probabilities

$$\mathrm{P}(d|f) = \frac{\exp\left(\theta \cdot \phi(d)\right)}{\sum_{d'} \exp\left(\theta \cdot \phi(d')\right)}$$

- We can query this distribution for:

Whole translations [Blunsom et al., '08]:

$$e_{\text{MAX-TRANS}} = \arg\max_{e} \sum_{d:\sigma_e(d)=e} \mathrm{P}(d|f)$$

N-gram overlap [Kumar and Byrne, '04]:

$$e_{\text{MBR}} = \arg\max_{e} \mathbb{E}\left[\text{BLEU}(e; \sigma_e(d))\right]$$

Consensus Decoding

- Derivation scores can be interpreted as probabilities

$$\mathrm{P}(d|f) = \frac{\exp\left(\theta \cdot \phi(d)\right)}{\sum_{d'} \exp\left(\theta \cdot \phi(d')\right)}$$

- We can query this distribution for:

Whole translations [Blunsom et al., '08]:

$$e_{\text{MAX-TRANS}} = \arg\max_{e} \sum_{d:\sigma_e(d)=e} \mathrm{P}(d|f)$$

N-gram overlap [Kumar and Byrne, '04]:

$$e_{\text{MBR}} = \arg\max_{e} \mathbb{E}\left[\text{BLEU}(e;\sigma_e(d))\right]$$

System Combination

- We often have multiple translation systems

System Combination

- We often have multiple translation systems

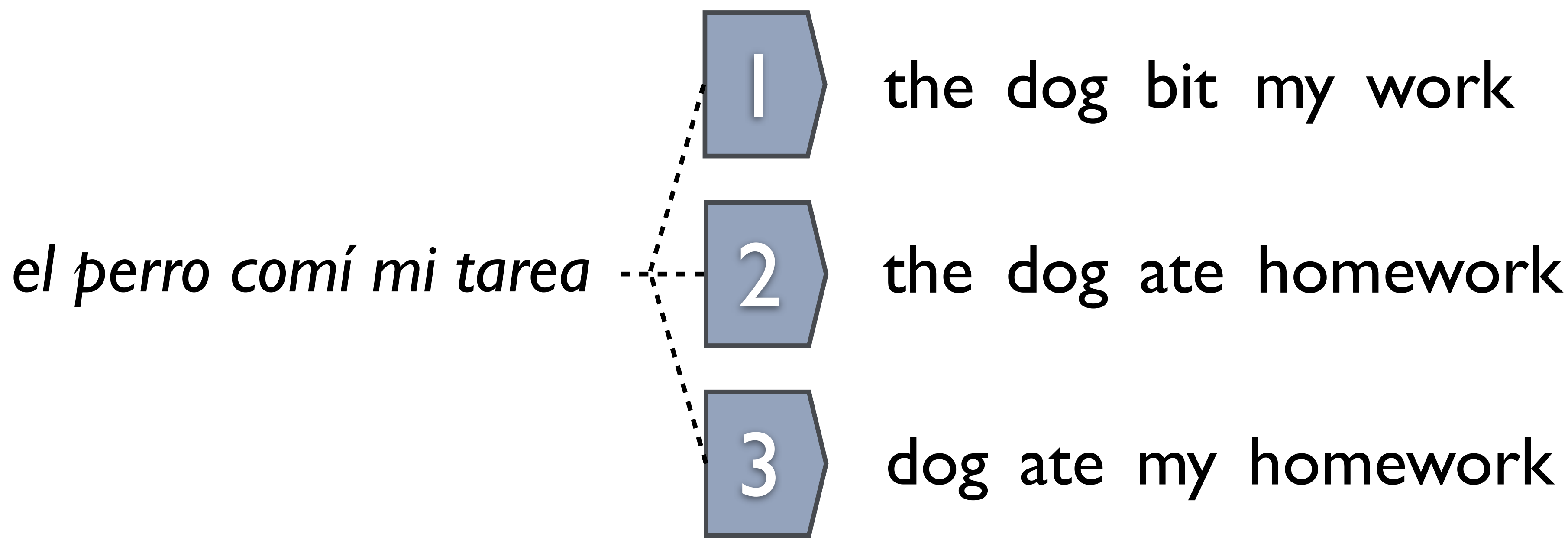
el perro comí mi tarea

System Combination

- We often have multiple translation systems

System Combination

- We often have multiple translation systems



System Combination

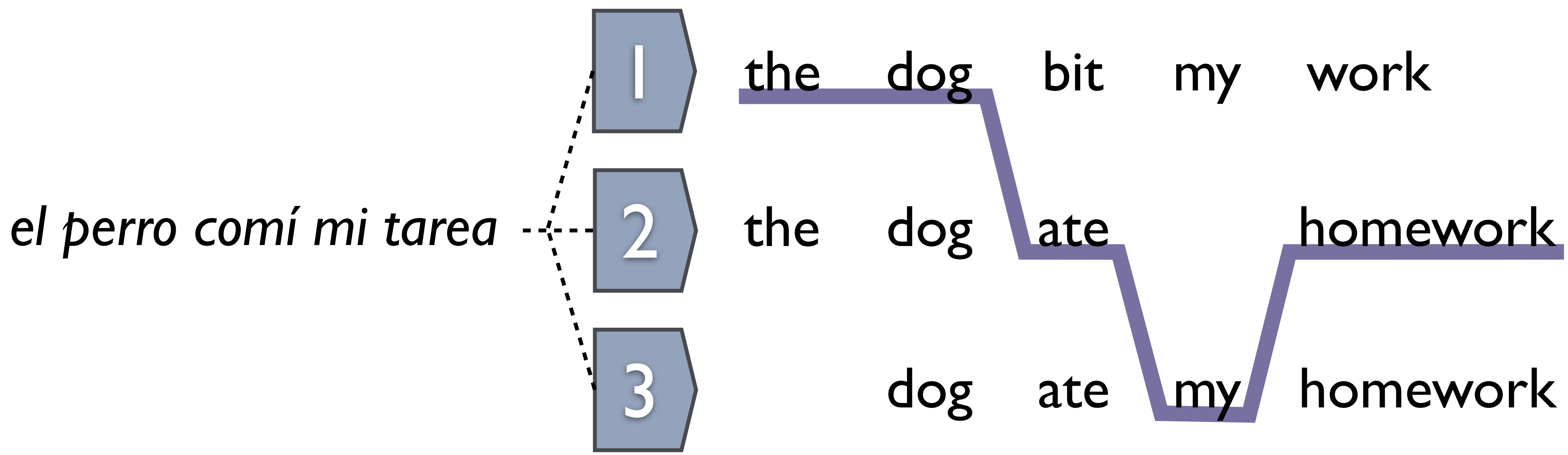
- We often have multiple translation systems

el perro comí mi tarea

1	the	dog	bit	my	work
2	the	dog	ate		homework
3		dog	ate	my	homework

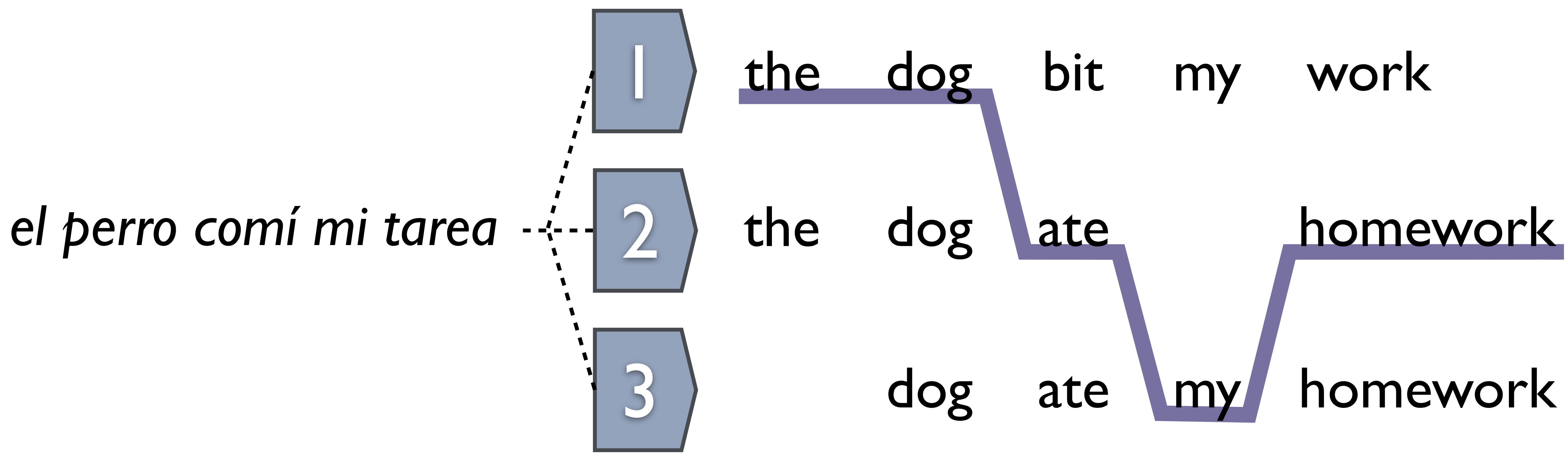
System Combination

- We often have multiple translation systems



System Combination

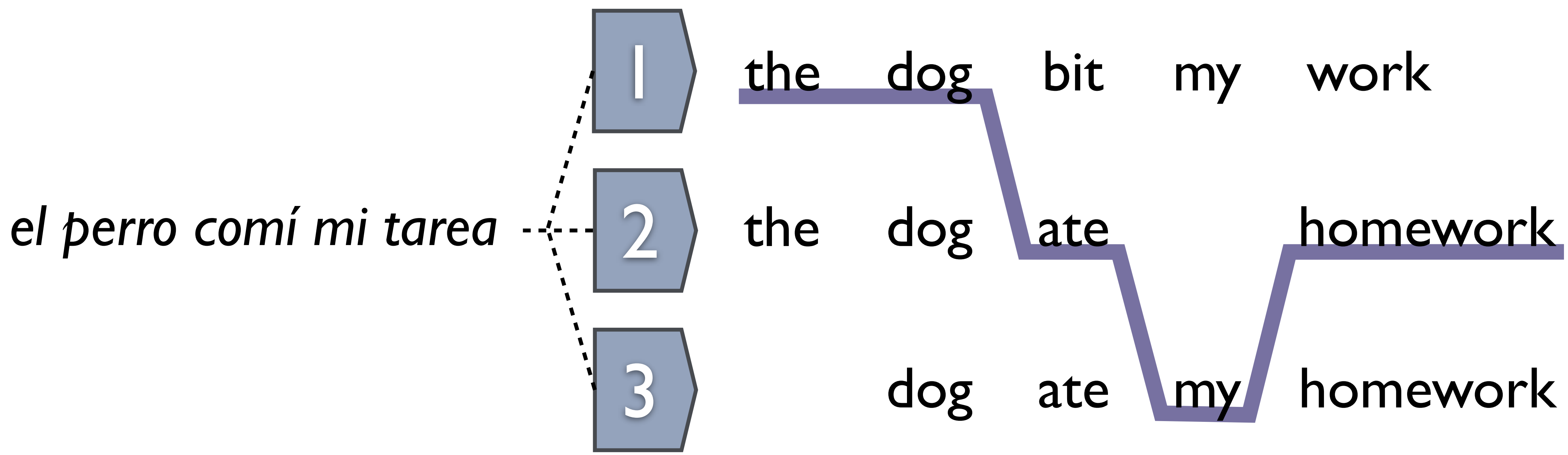
- We often have multiple translation systems



- Combiners assume little about systems

System Combination

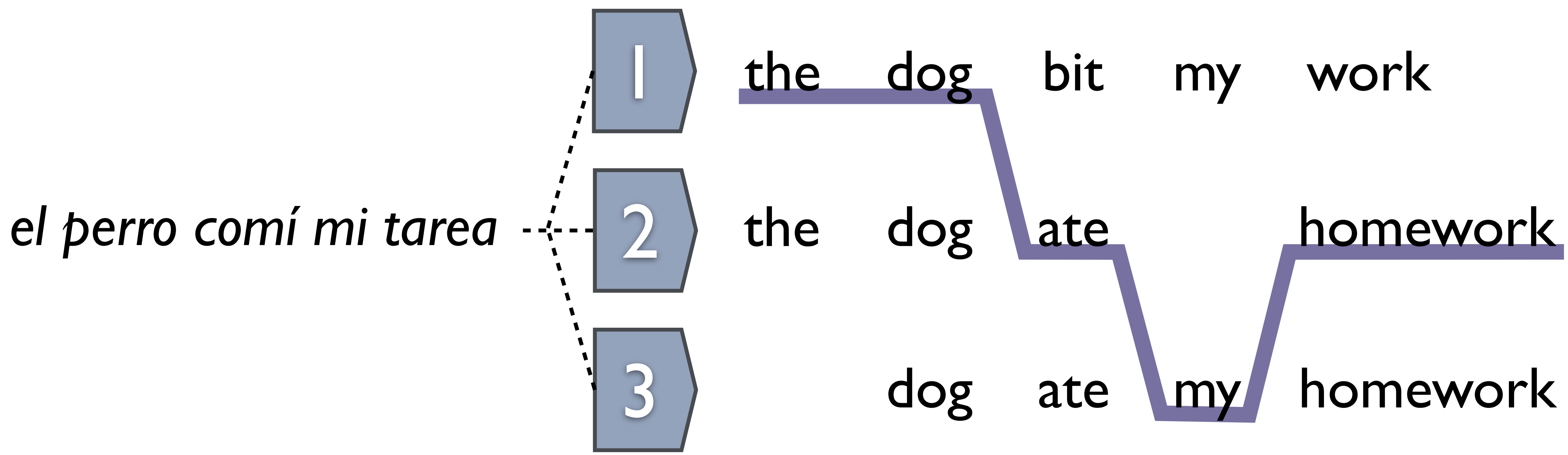
- We often have multiple translation systems



- Combiners assume little about systems
- Objectives similar to consensus decoding

System Combination

- We often have multiple translation systems



- Combiners assume little about systems
- Objectives similar to consensus decoding

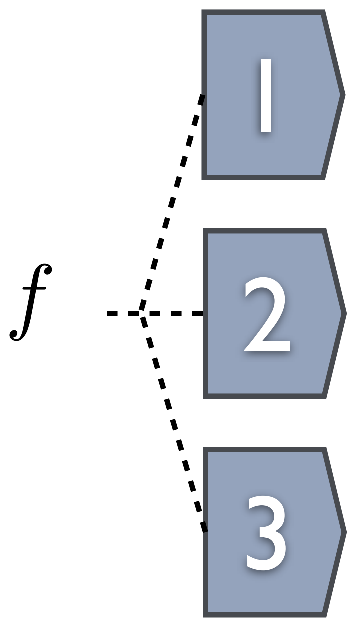
Model Combination

Goooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooogle

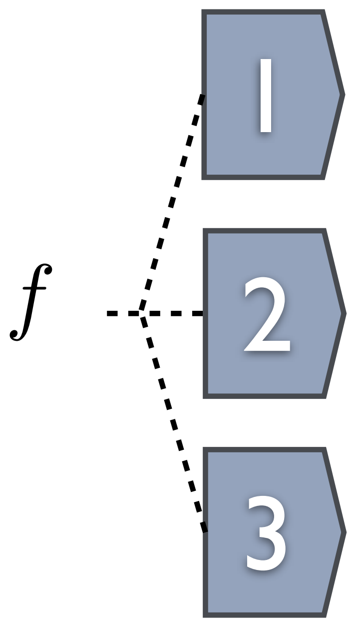
Model Combination

f

Model Combination



Model Combination

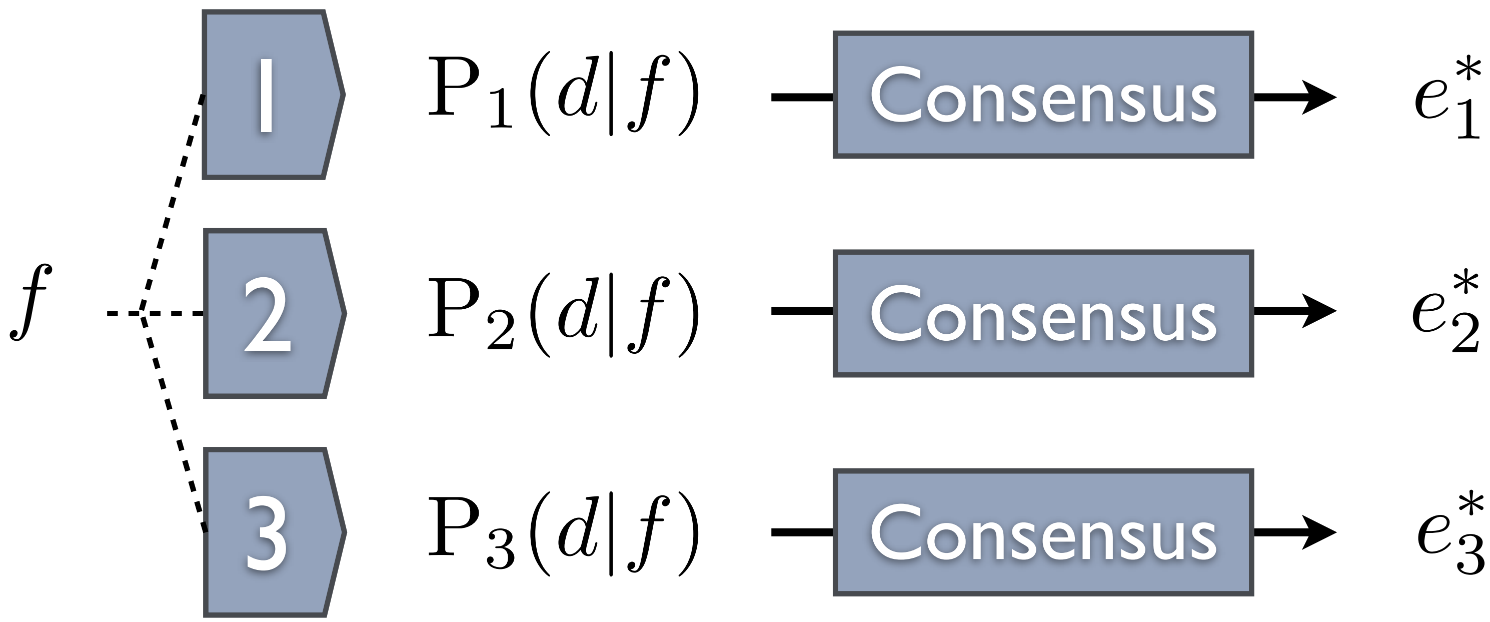


$$\mathrm{P}_1(d|f)$$

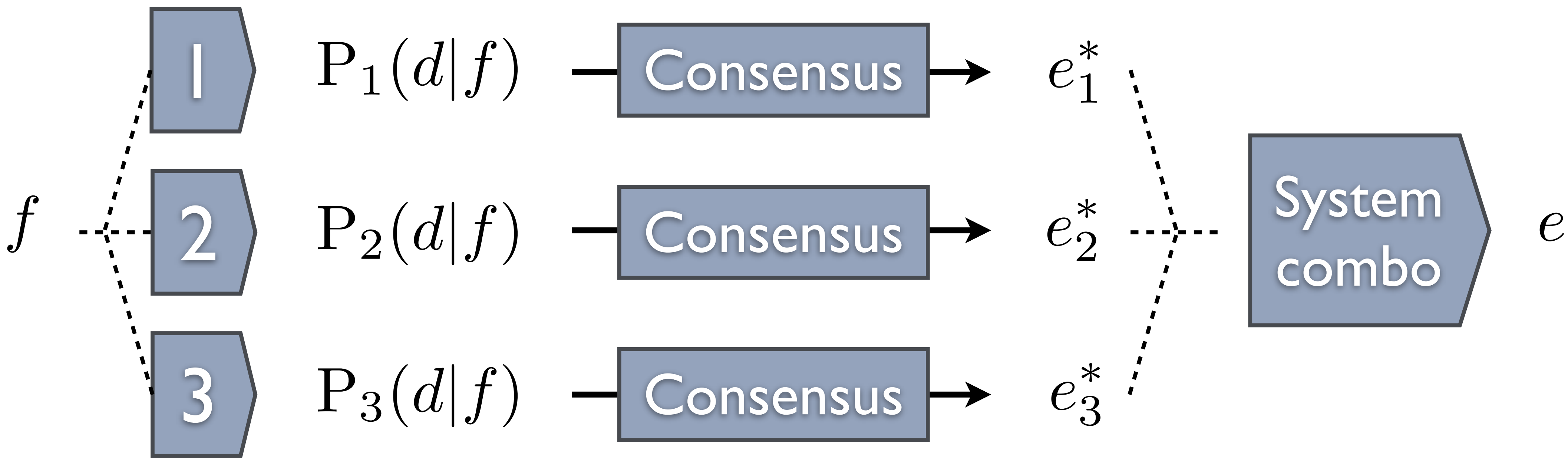
$$\mathrm{P}_2(d|f)$$

$$\mathrm{P}_3(d|f)$$

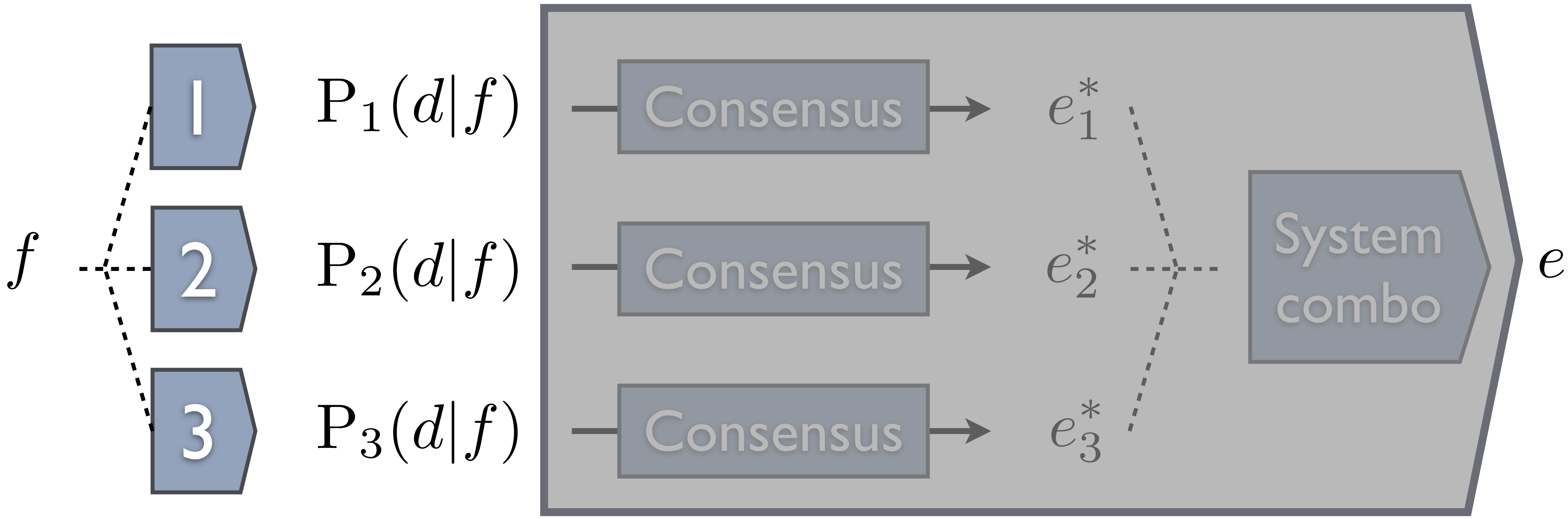
Model Combination



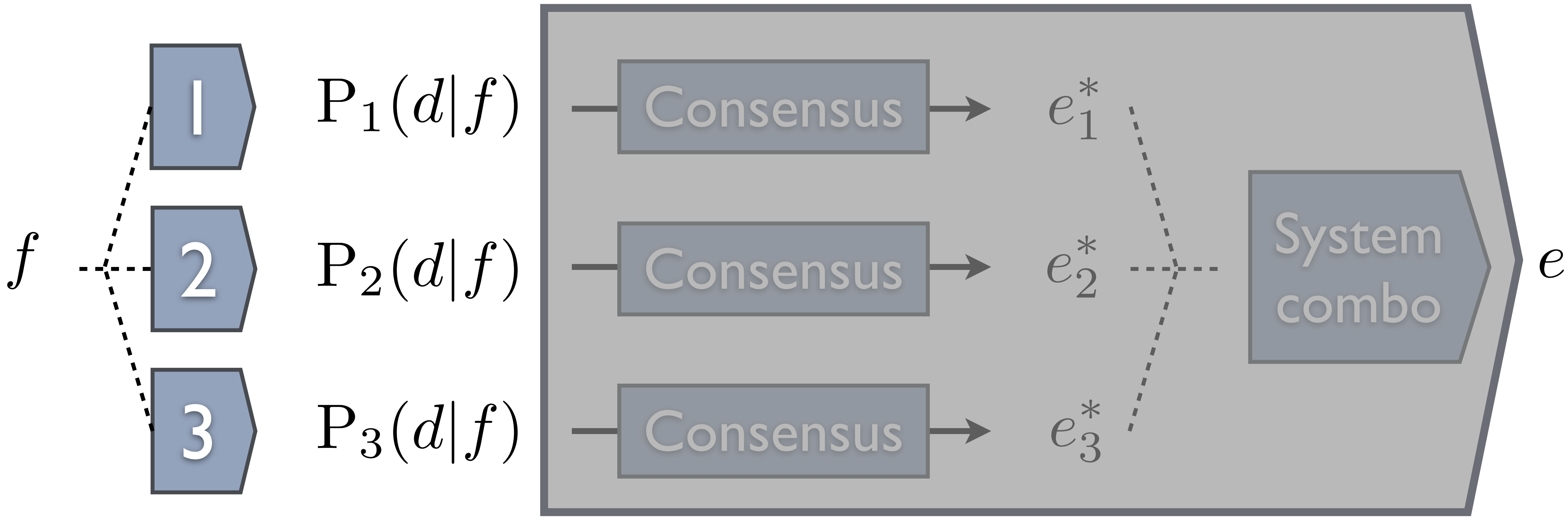
Model Combination



Model Combination

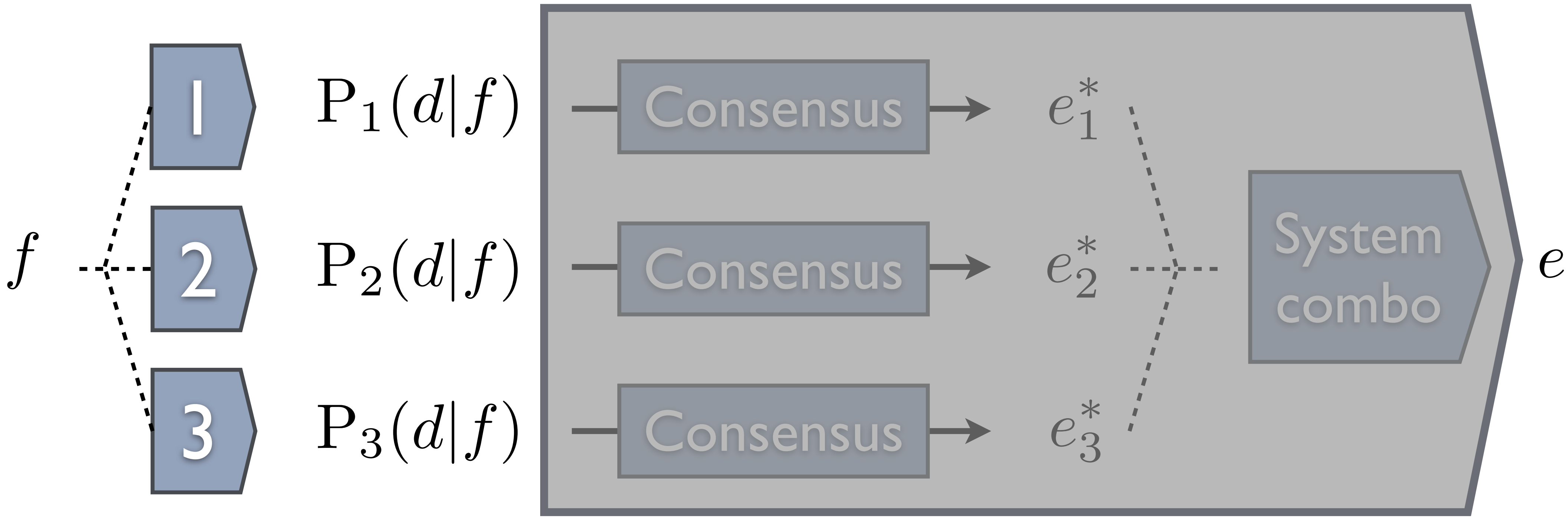


Model Combination



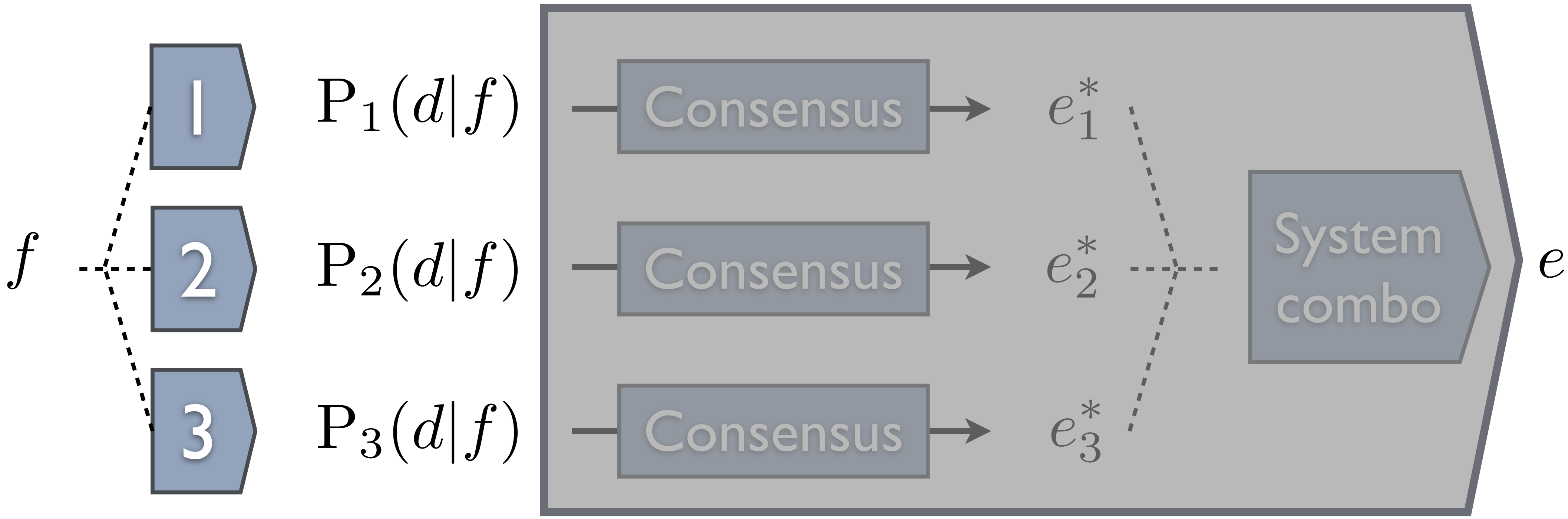
- Consensus decoding with multiple models

Model Combination



- Consensus decoding with multiple models
- Distribution-driven approach to system combination

Model Combination



- Consensus decoding with multiple models
- Distribution-driven approach to system combination
- Unifies consensus and combination objectives

Outline

Consensus decoding review

Our model combination technique

Comparison to system combination

Outline

Consensus decoding review

With new algorithms and results

Our model combination technique

Comparison to system combination

Forest-Based Consensus Decoding

Review

Gooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooogle

Forest-Based Consensus Decoding

Review

Gooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooogle

1 Build a forest that encodes the model posterior

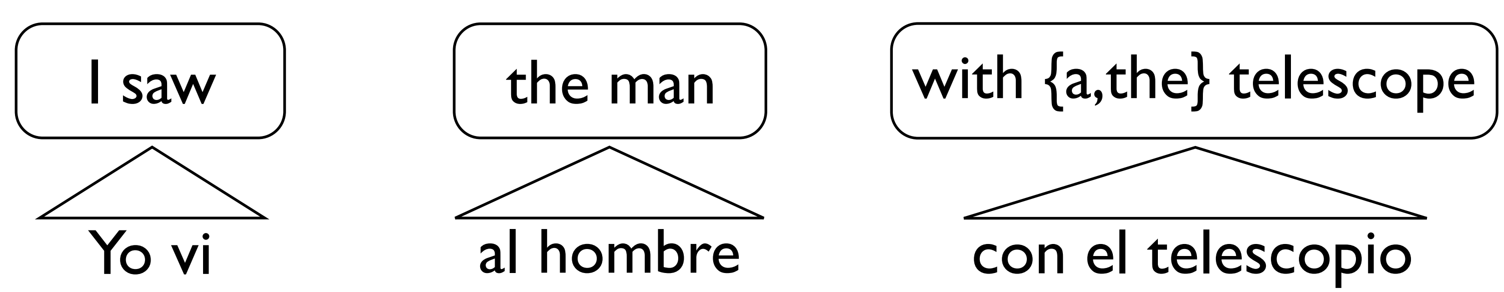
$$\mathrm{P}(d|f) =$$

Review

Forest-Based Consensus Decoding

1 Build a forest that encodes the model posterior

$$P(d|f) =$$

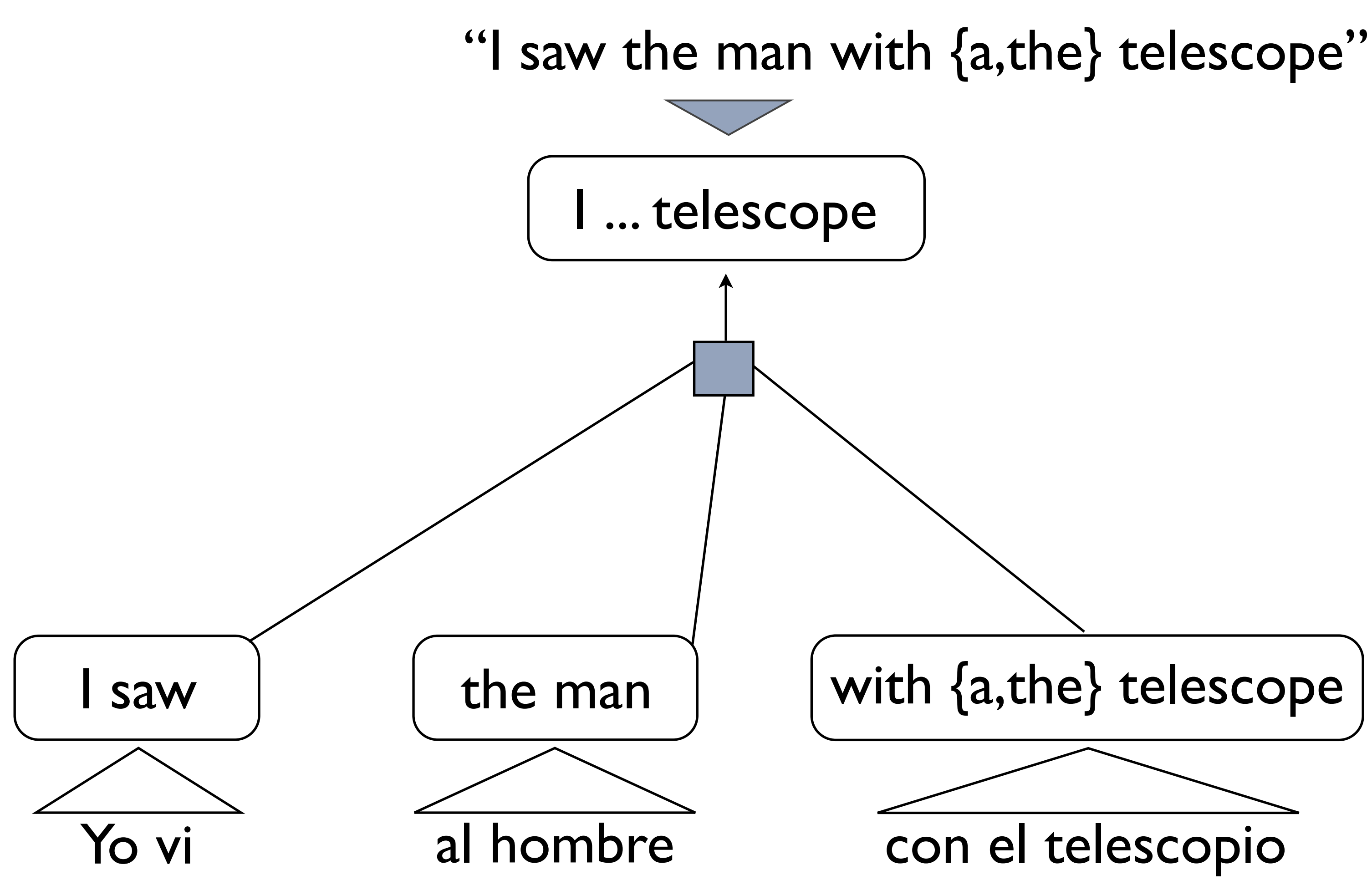


Forest-Based Consensus Decoding

Review

1 Build a forest that encodes the model posterior

$$\mathrm{P}(d|f) =$$



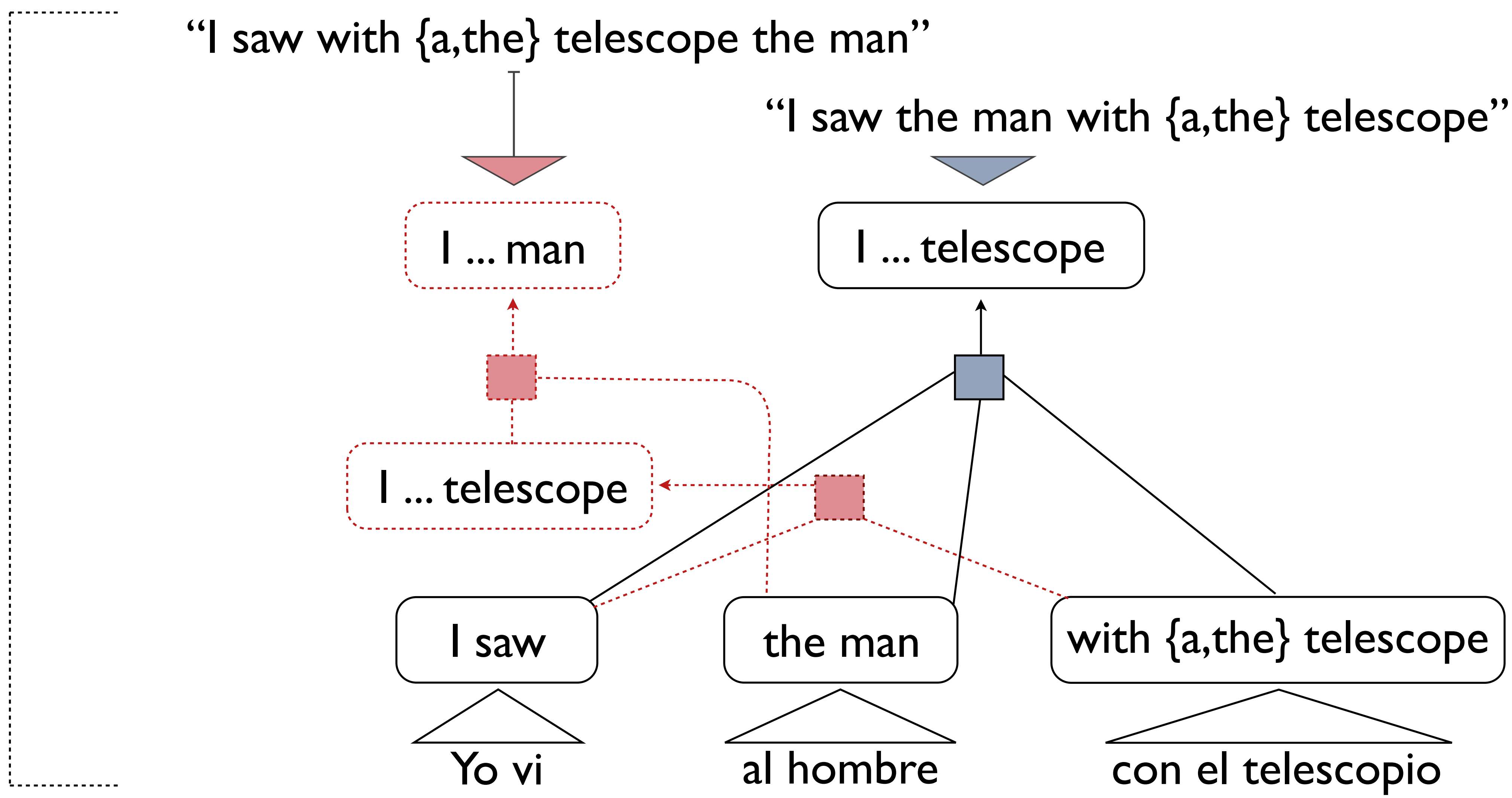
Review

Forest-Based Consensus Decoding

Gooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooogle

1 Build a forest that encodes the model posterior

$$P(d|f) =$$

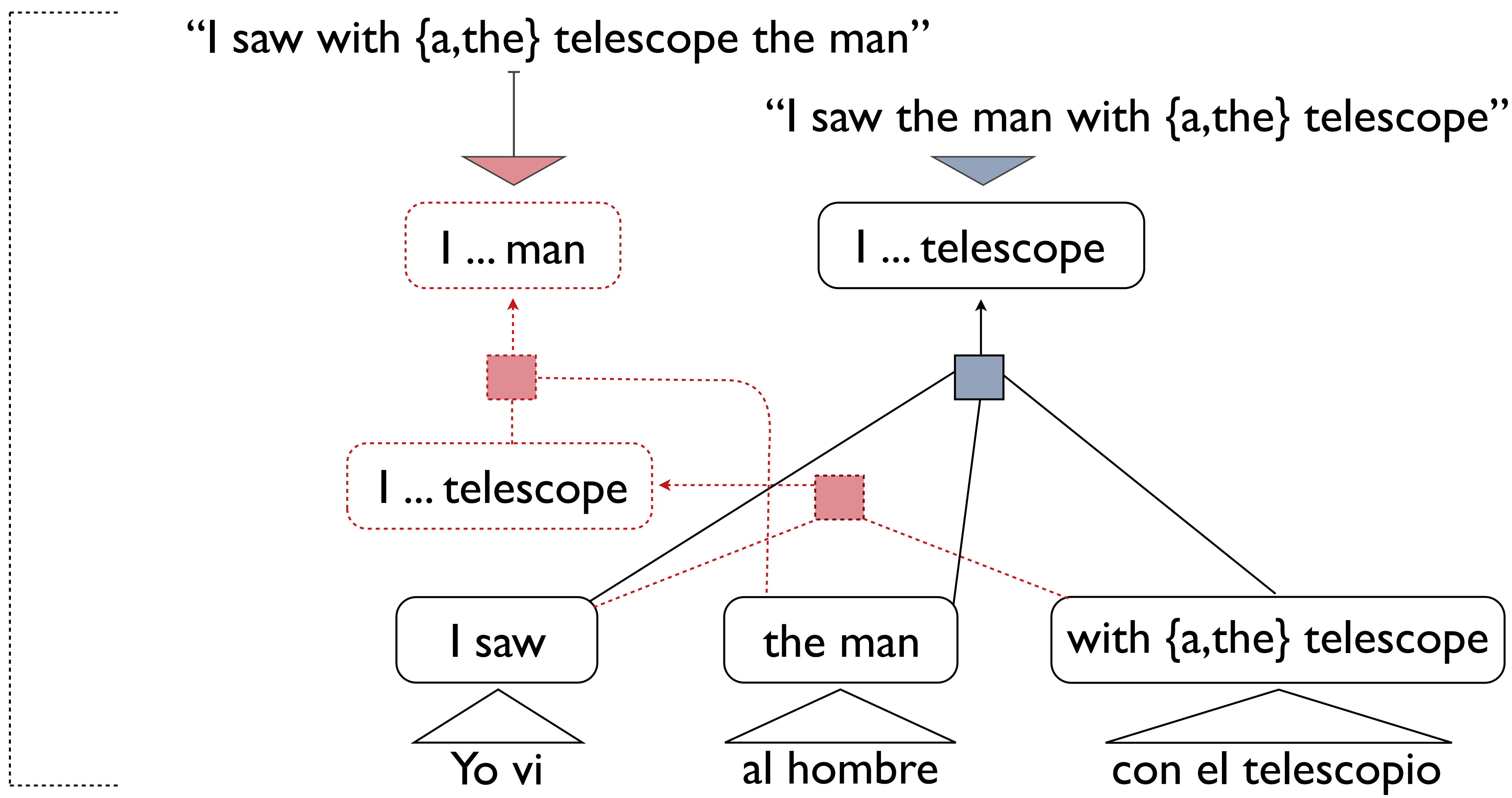


Review

Forest-Based Consensus Decoding

1. Build a forest that encodes the model posterior
2. Compute *n-gram statistics* from the posterior

$$P(d|f) =$$

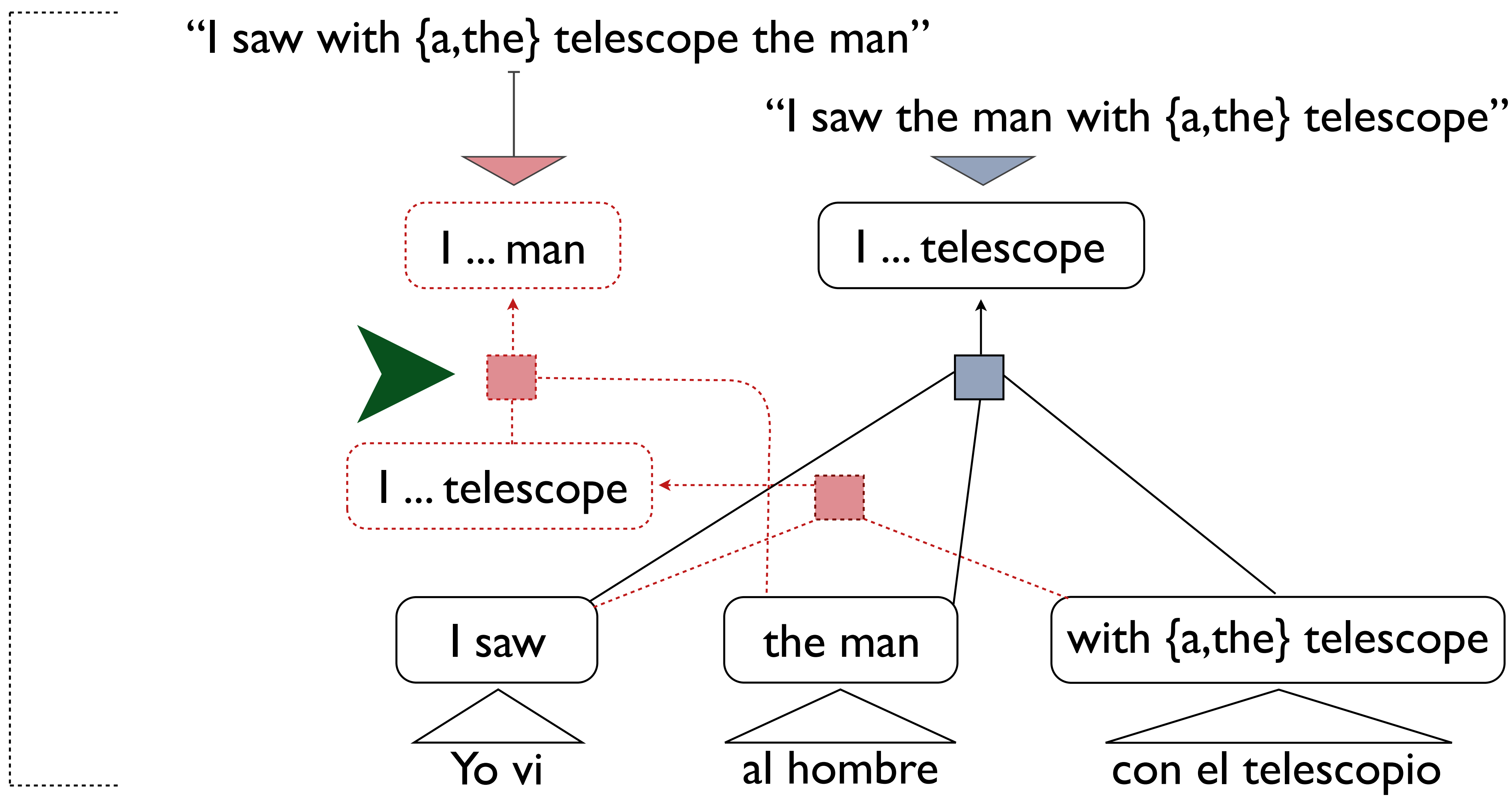


Review

Forest-Based Consensus Decoding

1. Build a forest that encodes the model posterior
2. Compute *n-gram statistics* from the posterior

$P(d|f) =$

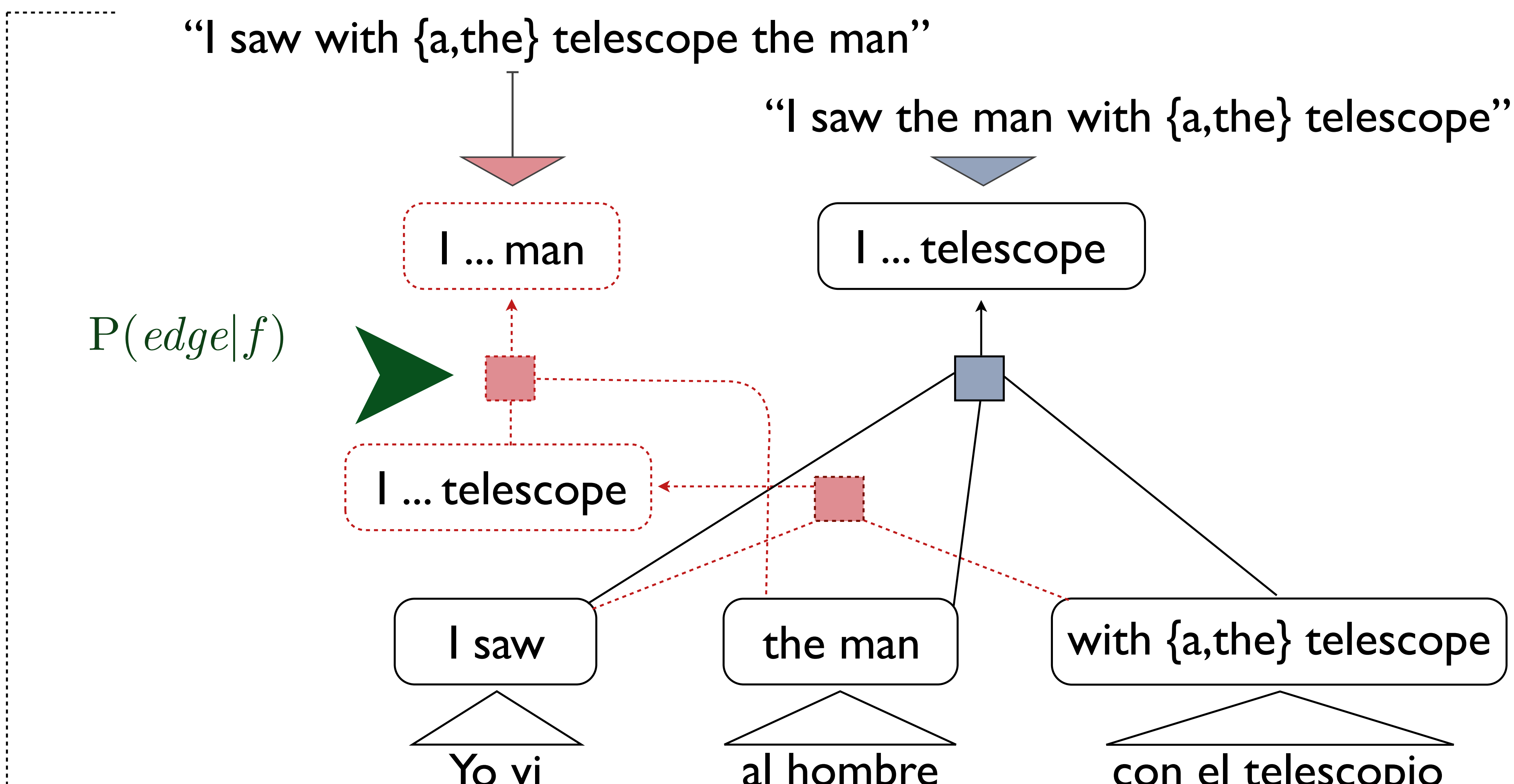


Review

Forest-Based Consensus Decoding

1. Build a forest that encodes the model posterior
2. Compute *n-gram statistics* from the posterior

$$P(d|f) =$$



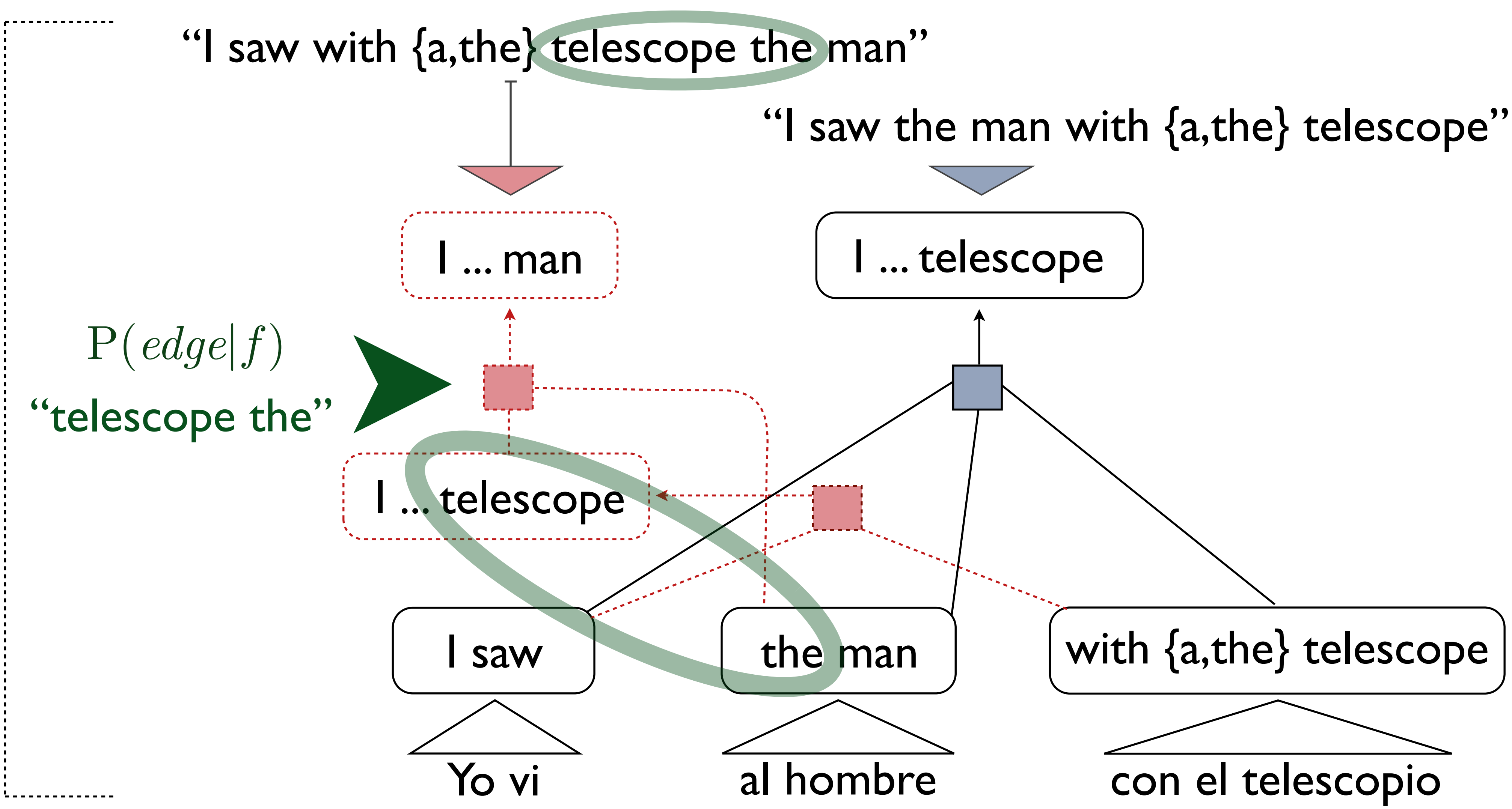
Forest-Based Consensus Decoding

Review

Gooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooogle

1. Build a forest that encodes the model posterior
2. Compute *n-gram statistics* from the posterior

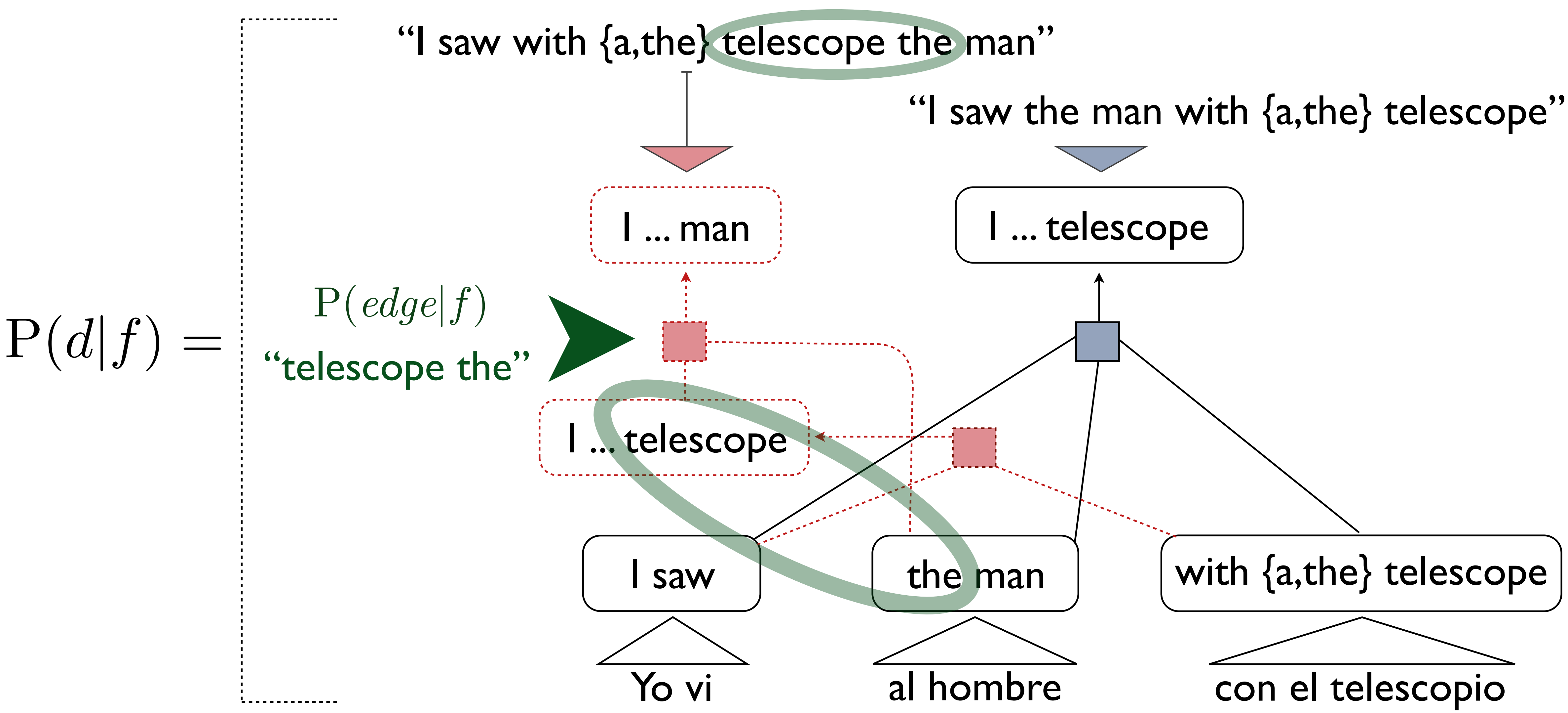
$$\mathrm{P}(d|f) =$$



Review

Forest-Based Consensus Decoding

1. Build a forest that encodes the model posterior
2. Compute *n-gram statistics* from the posterior
3. Optimize a *consensus objective* using these statistics



Types of Efficient Consensus Techniques

Review

Lattice Minimum Bayes-Risk Decoding [Tromble et al., '08]	

Review

Types of Efficient Consensus Techniques

Lattice Minimum Bayes-Risk Decoding [Tromble et al., ’08]	
Posteriors	Expected counts

N-gram Statistics

Review

Types of Efficient Consensus Techniques

Consensus Objective

	Posteriors	Expected counts
Learned		
Fixed	*Lattice Minimum Bayes-Risk Decoding* [Tromble et al., '08]	

N-gram Statistics

Types of Efficient Consensus Techniques

Review

Gooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooogle

Consensus Objective

	Posteriors	Expected counts
Learned	*Minimum Bayes-Risk Decoding for Hypergraphs* [Kumar et al., '09]	*Variational Decoding for Machine Translation* [Li et al., '09]
Fixed	*Lattice Minimum Bayes-Risk Decoding* [Tromble et al., '08]	*Fast Consensus Decoding* [DeNero et al., '09]

N-gram Statistics

Learned Objectives are Better than Fixed

Review

Gooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooogle

Learned Objectives are Better than Fixed

Review

Goooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooogle

$$\mathrm{C}(d) = \sum_{n=1}^{4} w^{(n)} \sum_{g \in n\text{-grams}} c(g,d) \cdot \mathrm{P}(g|f) + w^{(\ell)} |\sigma_e(d)| + w^{(b)} \theta \cdot \phi(d)$$

Review

Learned Objectives are Better than Fixed

N-gram statistics

$$\mathrm{C}(d) = \sum_{n=1}^{4} w^{(n)} \sum_{g \in n\text{-grams}} c(g,d) \cdot \mathrm{P}(g|f) + w^{(\ell)} |\sigma_e(d)| + w^{(b)} \theta \cdot \phi(d)$$

Review

Learned Objectives are Better than Fixed

Gooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooogle

$$\mathrm{C}(d) = \underbrace{\sum_{n=1}^{4} w^{(n)} \sum_{g \in n\text{-grams}} c(g,d) \cdot \mathrm{P}(g|f)}_{\text{N-gram statistics}} + \underbrace{w^{(\ell)} |\sigma_e(d)|}_{\text{Length}} + w^{(b)} \theta \cdot \phi(d)$$

Review

Learned Objectives are Better than Fixed

N-gram statistics — Length — Base model

$$\mathrm{C}(d) = \sum_{n=1}^{4} w^{(n)} \sum_{g \in n\text{-grams}} c(g,d) \cdot \mathrm{P}(g|f) + w^{(\ell)} |\sigma_e(d)| + w^{(b)} \theta \cdot \phi(d)$$

Review

Learned Objectives are Better than Fixed

N-gram statistics | Length | Base model

$$\mathrm{C}(d) = \sum_{n=1}^{4} w^{(n)} \sum_{g \in n\text{-grams}} c(g,d) \cdot \mathrm{P}(g|f) + w^{(\ell)} |\sigma_e(d)| + w^{(b)} \theta \cdot \phi(d)$$

Objective parameters

Review

Learned Objectives are Better than Fixed

Goooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooogle

N-gram statistics   Length   Base model

$$\mathrm{C}(d) = \sum_{n=1}^{4} w^{(n)} \sum_{g \in n\text{-grams}} c(g,d) \cdot \mathrm{P}(g|f) + w^{(\ell)} |\sigma_e(d)| + w^{(b)} \theta \cdot \phi(d)$$

Objective parameters

Fixed: Choose w such that $C_w(d) \approx \mathbb{E}\left[\mathrm{BLEU}(d)\right]$

Learned Objectives are Better than Fixed

Review

Gooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooogle

N-gram statistics Length Base model

$$\mathrm{C}(d) = \sum_{n=1}^{4} w^{(n)} \sum_{g \in n\text{-grams}} c(g, d) \cdot \mathrm{P}(g|f) + w^{(\ell)} |\sigma_e(d)| + w^{(b)} \theta \cdot \phi(d)$$

Objective parameters

Fixed: Choose w such that $C_w(d) \approx \mathbb{E}\left[\mathrm{BLEU}(d)\right]$

Learned: Choose w to ***maximize*** $\mathrm{BLEU}\left(\left\{\arg\max_d C_w(d)\right\}; \mathbf{e}\right)$

[Kumar et al., '09]

Learned Objectives are Better than Fixed

Review

Goooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooogle

N-gram statistics — Length — Base model

$$C(d) = \sum_{n=1}^{4} w^{(n)} \sum_{g \in n\text{-grams}} c(g,d) \cdot P(g|f) + w^{(\ell)} |\sigma_e(d)| + w^{(b)} \theta \cdot \phi(d)$$

Objective parameters

Fixed: Choose w such that $C_w(d) \approx \mathbb{E}\left[\mathrm{BLEU}(d)\right]$

References

Learned: Choose w to ***maximize*** $\mathrm{BLEU}\left(\left\{\arg\max_d C_w(d)\right\}; \mathbf{e}\right)$

[Kumar et al., '09]

Review

Learned Objectives are Better than Fixed

$$\mathrm{C}(d) = \overbrace{\sum_{n=1}^{4} w^{(n)} \sum_{g \in n\text{-grams}} c(g,d) \cdot \mathrm{P}(g|f)}^{\text{N-gram statistics}} + \overbrace{w^{(\ell)} |\sigma_e(d)|}^{\text{Length}} + \overbrace{w^{(b)} \theta \cdot \phi(d)}^{\text{Base model}}$$

Objective parameters

Fixed: Choose w such that $C_w(d) \approx \mathbb{E}\left[\mathrm{BLEU}(d)\right]$

References

Learned: Choose w to ***maximize*** $\mathrm{BLEU}\left(\left\{\arg\max_{d} C_w(d)\right\}; \mathbf{e}\right)$

[Kumar et al., '09]

Consensus performance versus max-derivation decoding (39 pairs)

Increased test set BLEU by ≥0.2

Decreased test set BLEU by ≥0.2

Review

Learned Objectives are Better than Fixed

N-gram statistics | Length | Base model

$$\mathrm{C}(d) = \sum_{n=1}^{4} w^{(n)} \sum_{g \in n\text{-grams}} c(g,d) \cdot \mathrm{P}(g|f) + w^{(\ell)} |\sigma_e(d)| + w^{(b)} \theta \cdot \phi(d)$$

Objective parameters

Fixed: Choose w such that $C_w(d) \approx \mathbb{E}\left[\mathrm{BLEU}(d)\right]$

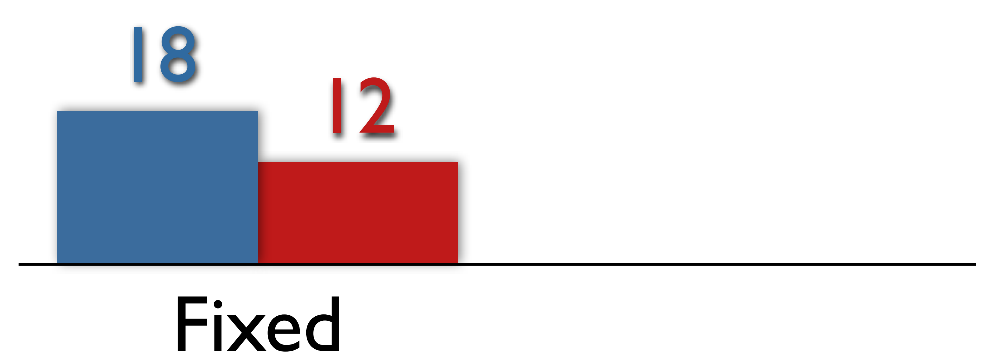
References

Learned: Choose w to ***maximize*** $\mathrm{BLEU}\left(\left\{\arg\max_d C_w(d)\right\}; \mathbf{e}\right)$

[Kumar et al., '09]

Consensus performance versus
max-derivation decoding (39 pairs)

- Increased test set BLEU by ≥0.2
- Decreased test set BLEU by ≥0.2



Review

Learned Objectives are Better than Fixed

N-gram statistics | Length | Base model

$$\mathrm{C}(d) = \sum_{n=1}^{4} w^{(n)} \sum_{g \in n\text{-grams}} c(g,d) \cdot \mathrm{P}(g|f) + w^{(\ell)} |\sigma_e(d)| + w^{(b)} \theta \cdot \phi(d)$$

Objective parameters

Fixed: Choose w such that $C_w(d) \approx \mathbb{E}\left[\mathrm{BLEU}(d)\right]$

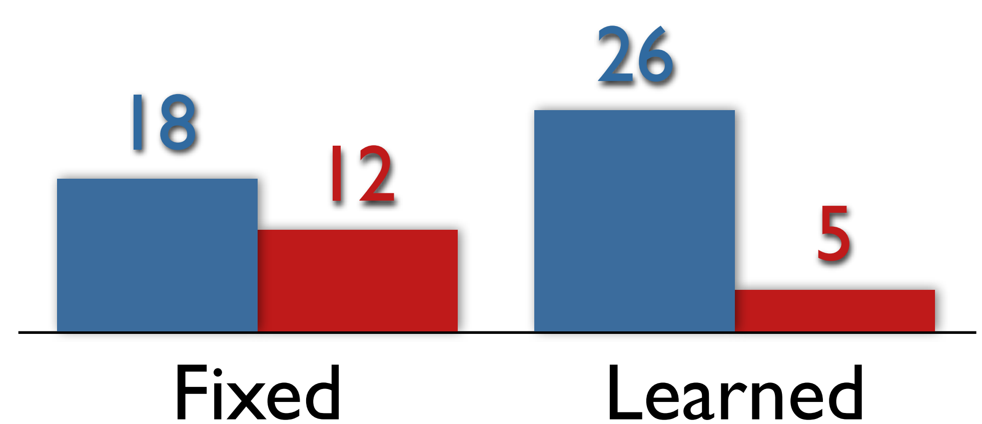
References

Learned: Choose w to ***maximize*** $\mathrm{BLEU}\left(\left\{\arg\max_d C_w(d)\right\}; \mathbf{e}\right)$

[Kumar et al., '09]

Consensus performance versus max-derivation decoding (39 pairs)

- Increased test set BLEU by ≥0.2
- Decreased test set BLEU by ≥0.2



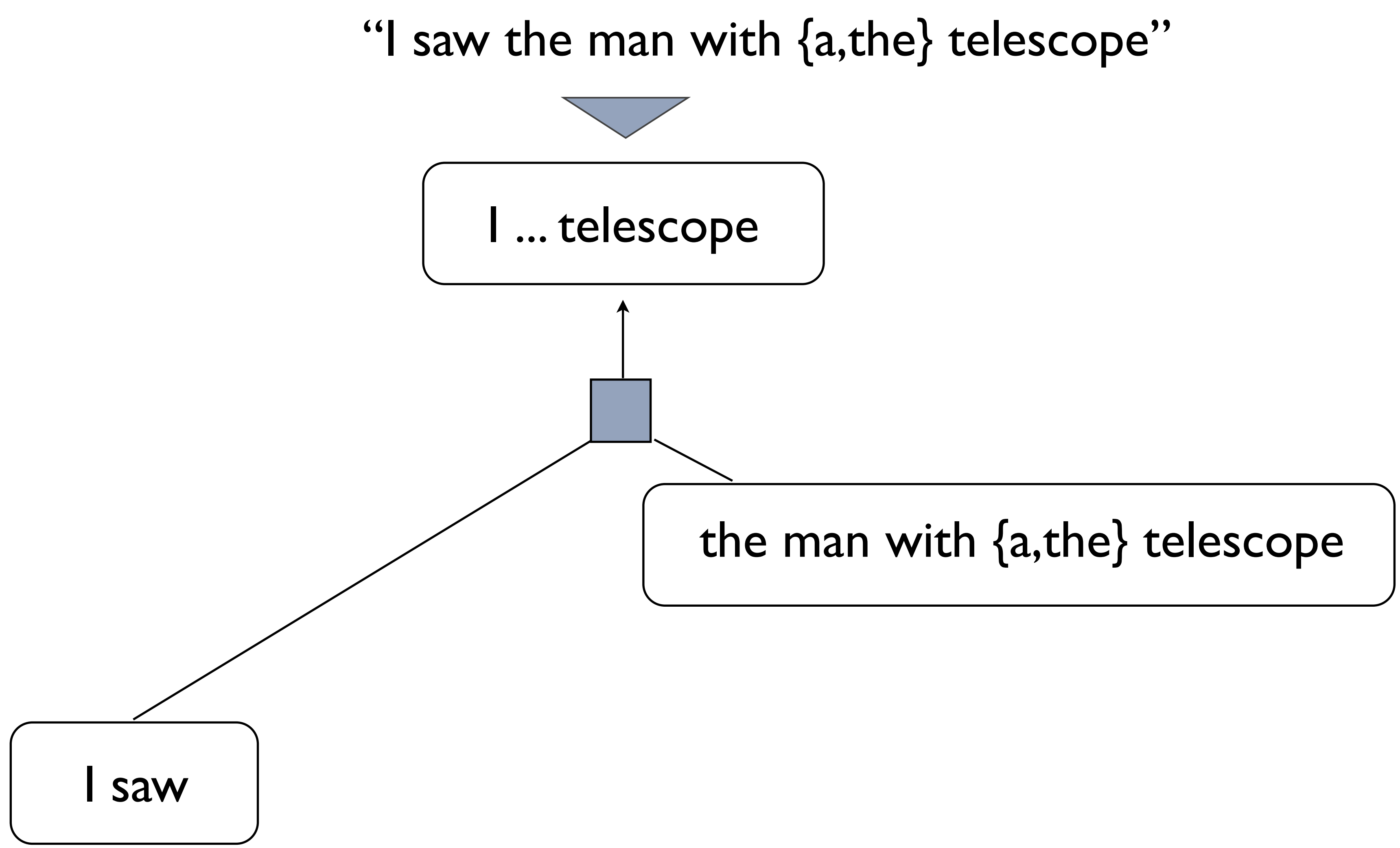
N-gram Posteriors can also be Computed Quickly

N-gram Posteriors can also be Computed Quickly

$$\mathrm{P}(g|f) = 1.0 - \mathrm{P}(\bar{g}|f)$$

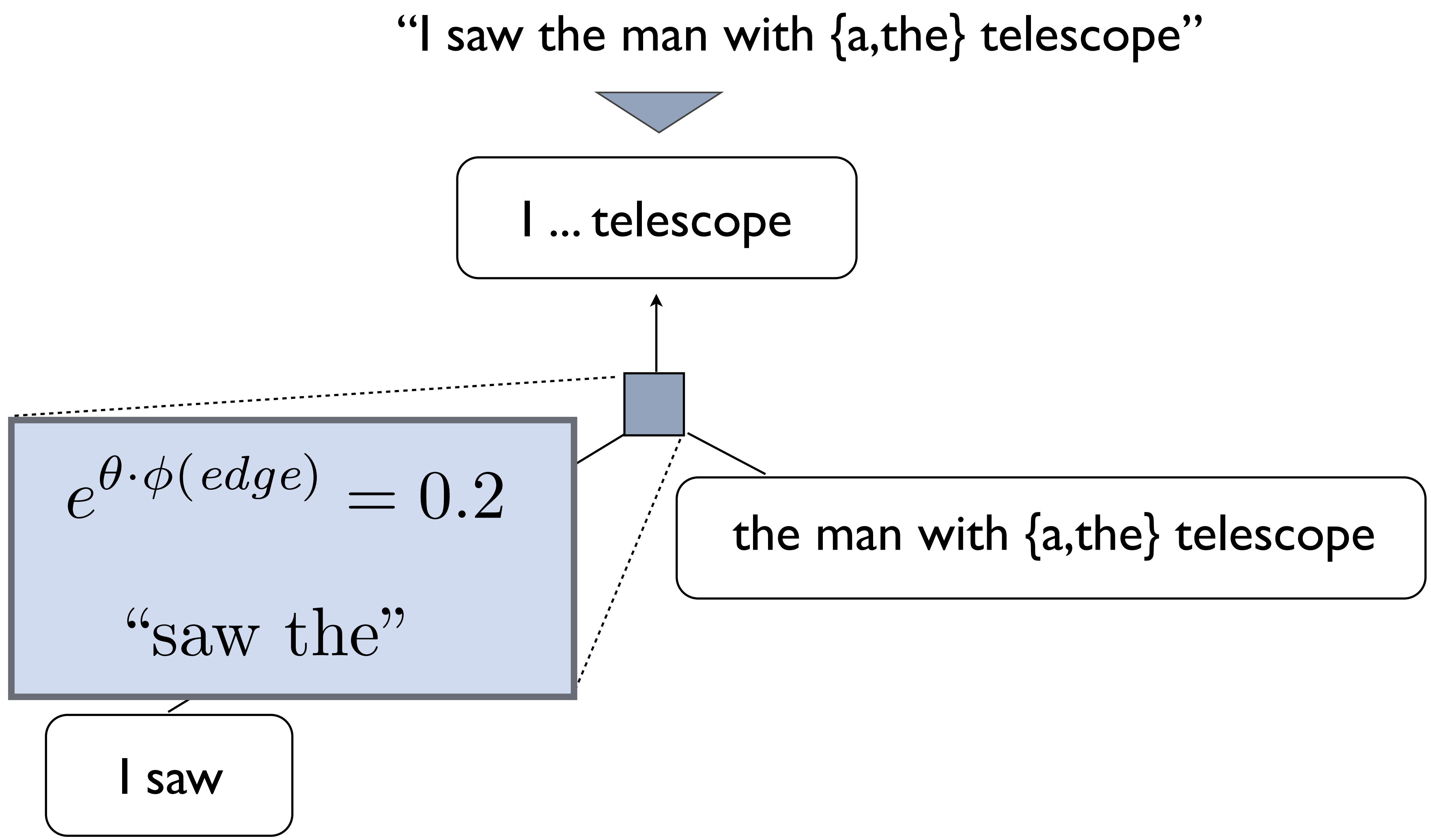
N-gram Posteriors can also be Computed Quickly

$$\mathrm{P}(g|f) = 1.0 - \mathrm{P}(\bar{g}|f)$$



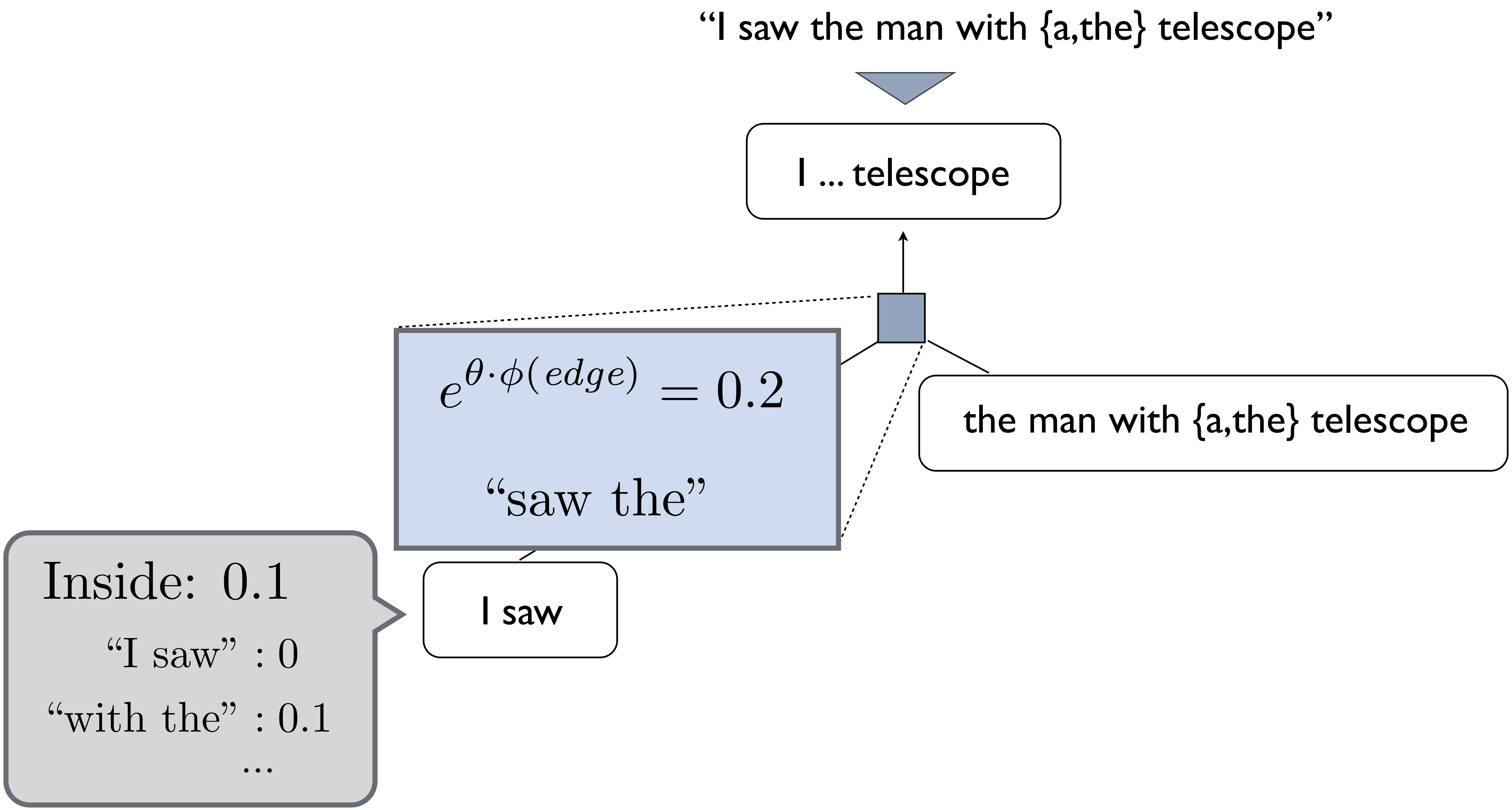
N-gram Posteriors can also be Computed Quickly

$$\mathrm{P}(g|f) = 1.0 - \mathrm{P}(\bar{g}|f)$$



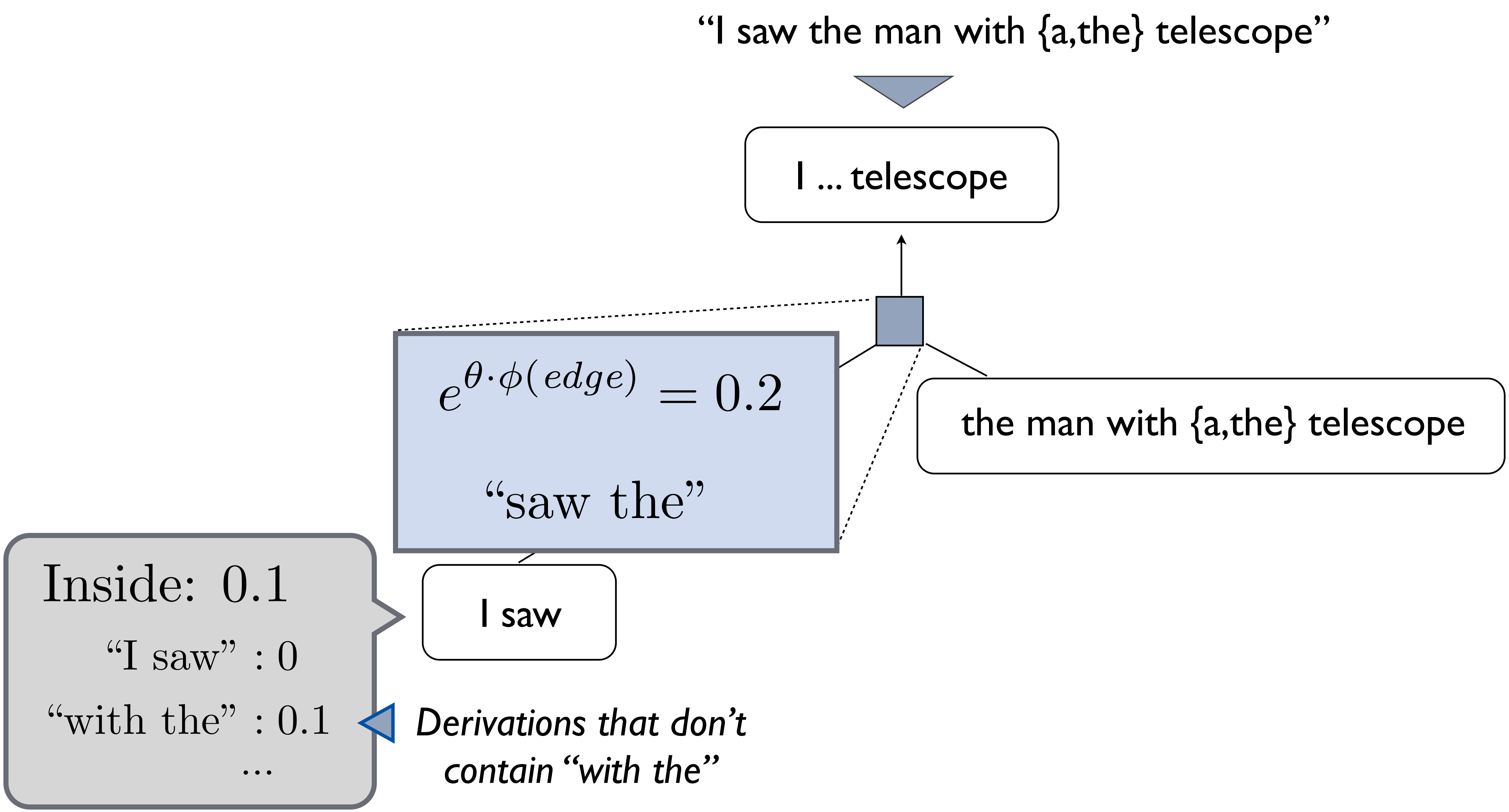
N-gram Posteriors can also be Computed Quickly

$$\mathrm{P}(g|f) = 1.0 - \mathrm{P}(\bar{g}|f)$$



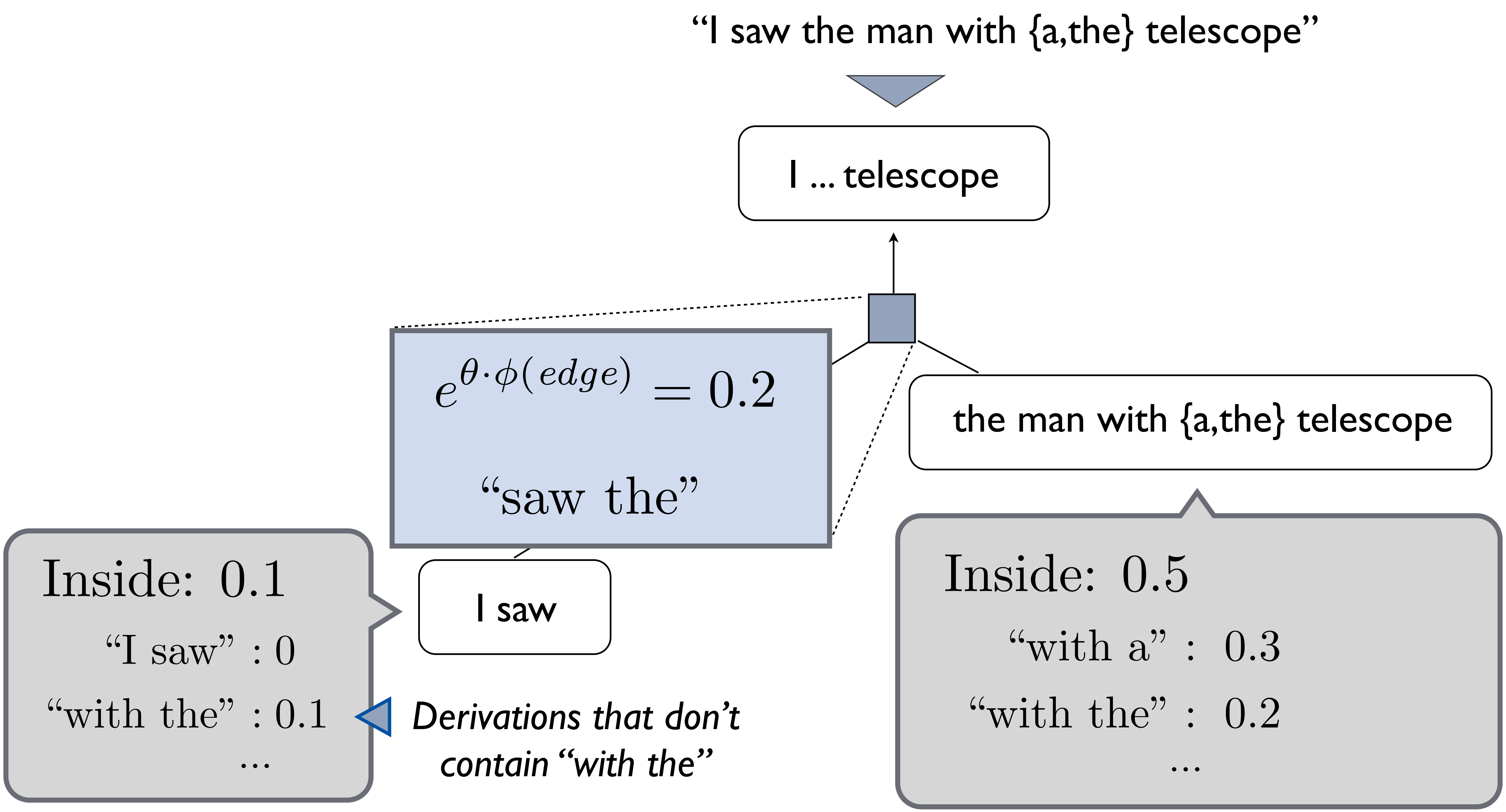
N-gram Posteriors can also be Computed Quickly

$$\mathrm{P}(g|f) = 1.0 - \mathrm{P}(\bar{g}|f)$$



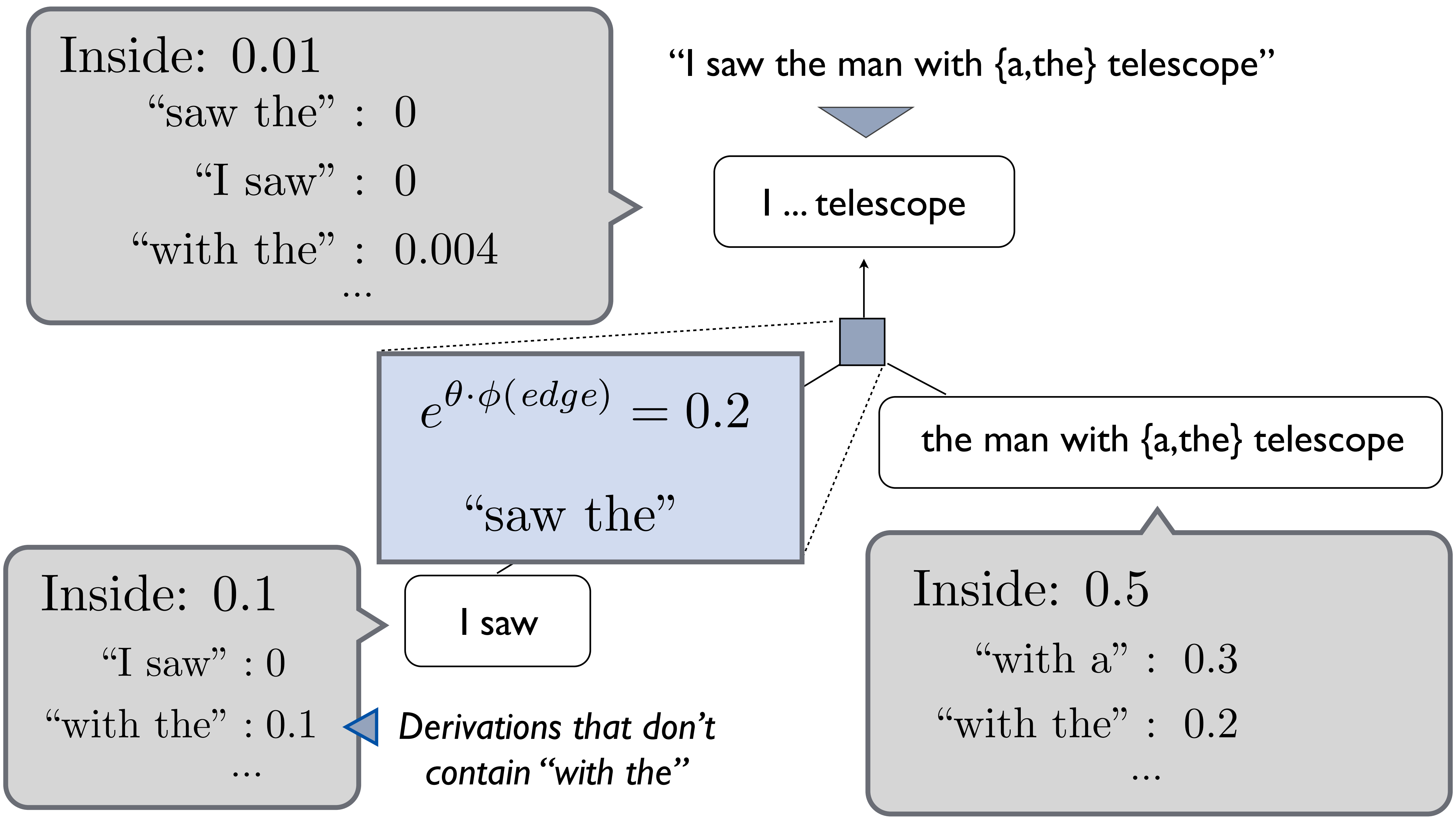
N-gram Posteriors can also be Computed Quickly

$$\mathrm{P}(g|f) = 1.0 - \mathrm{P}(\bar{g}|f)$$



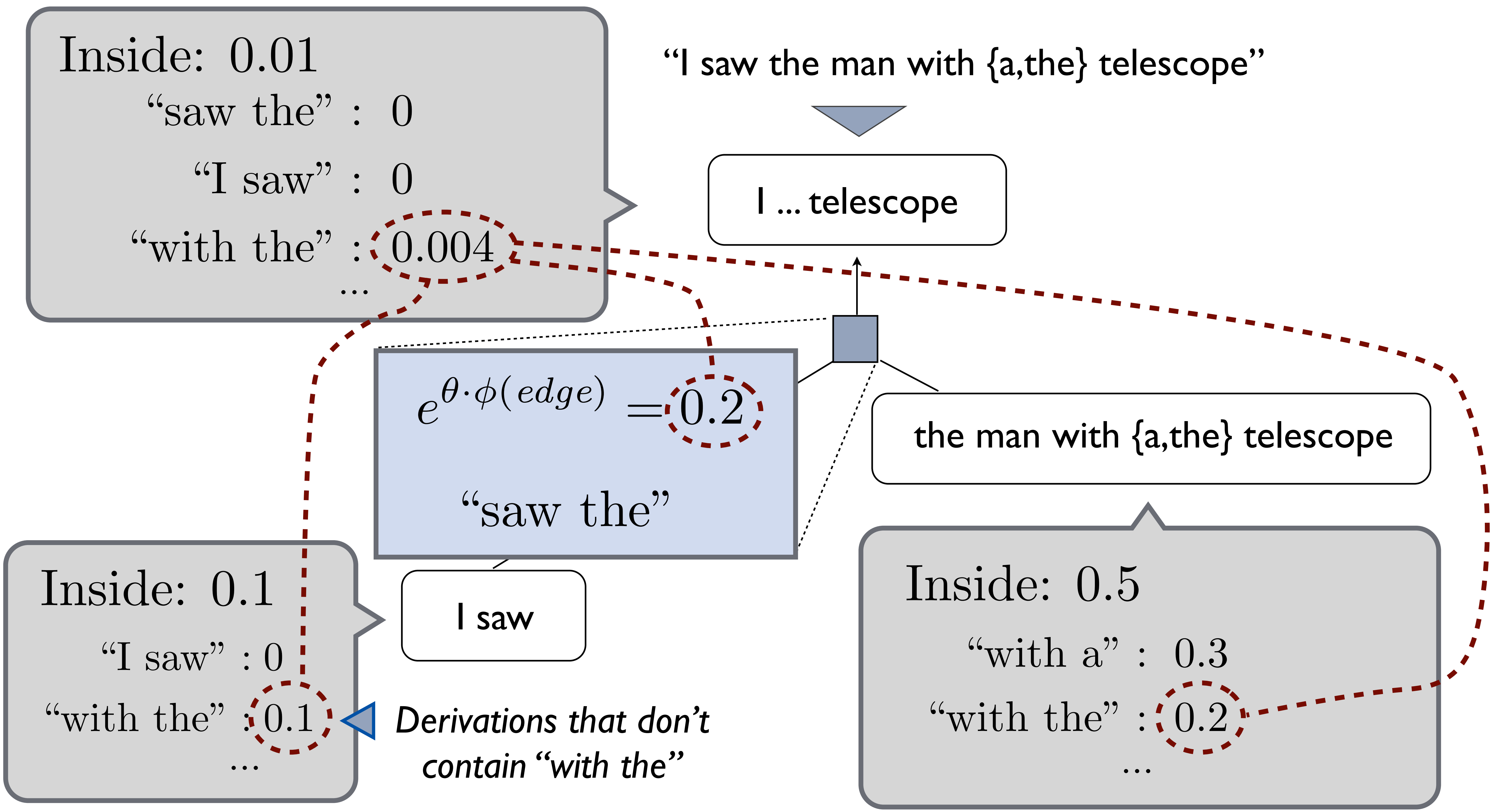
N-gram Posteriors can also be Computed Quickly

$$\mathrm{P}(g|f) = 1.0 - \mathrm{P}(\bar{g}|f)$$



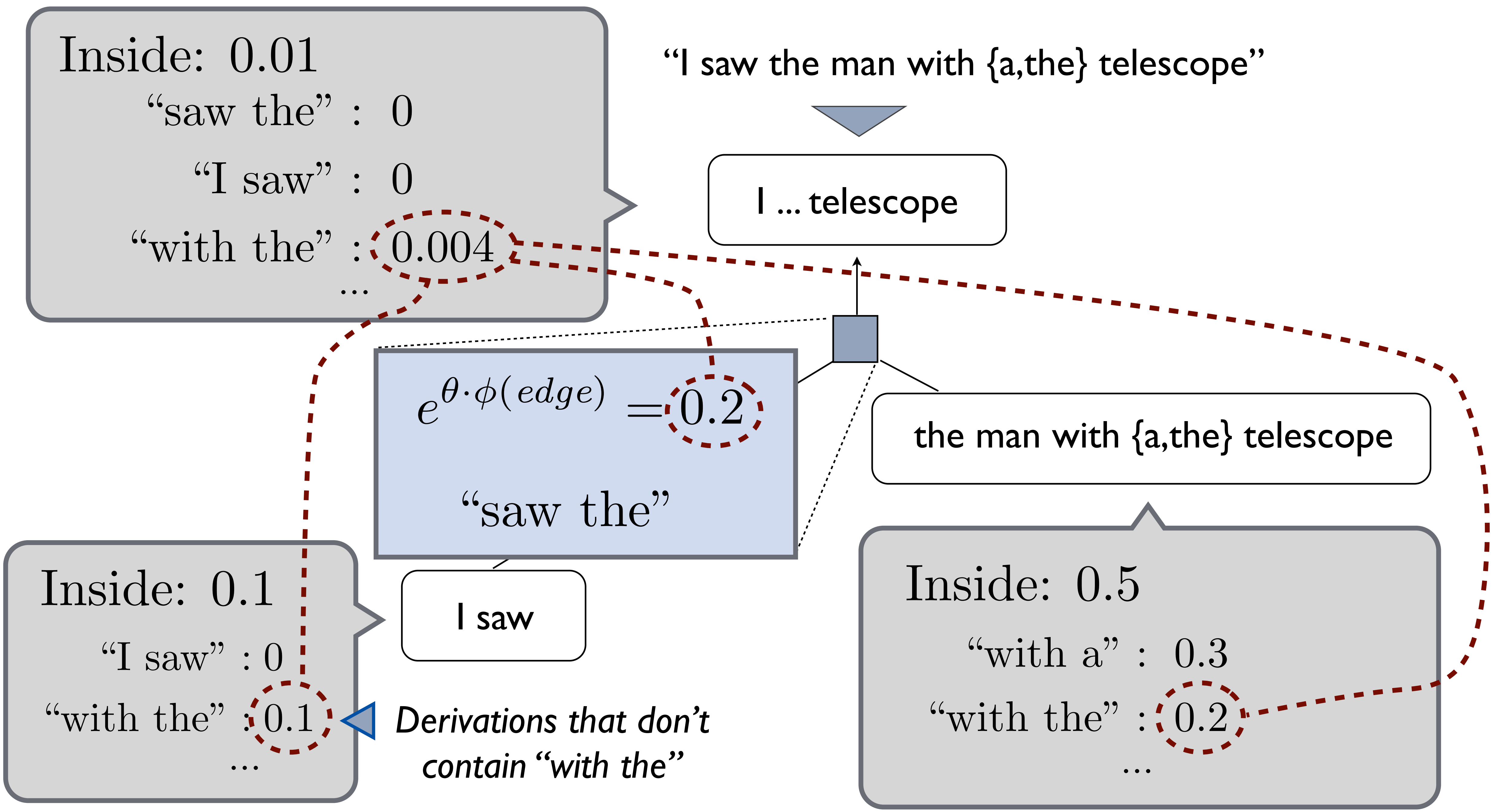
N-gram Posteriors can also be Computed Quickly

$$\mathrm{P}(g|f) = 1.0 - \mathrm{P}(\bar{g}|f)$$



N-gram Posteriors can also be Computed Quickly

$$\mathrm{P}(g|f) = 1.0 - \mathrm{P}(\bar{g}|f)$$



Audience challenge: What semiring computes n-gram posteriors?

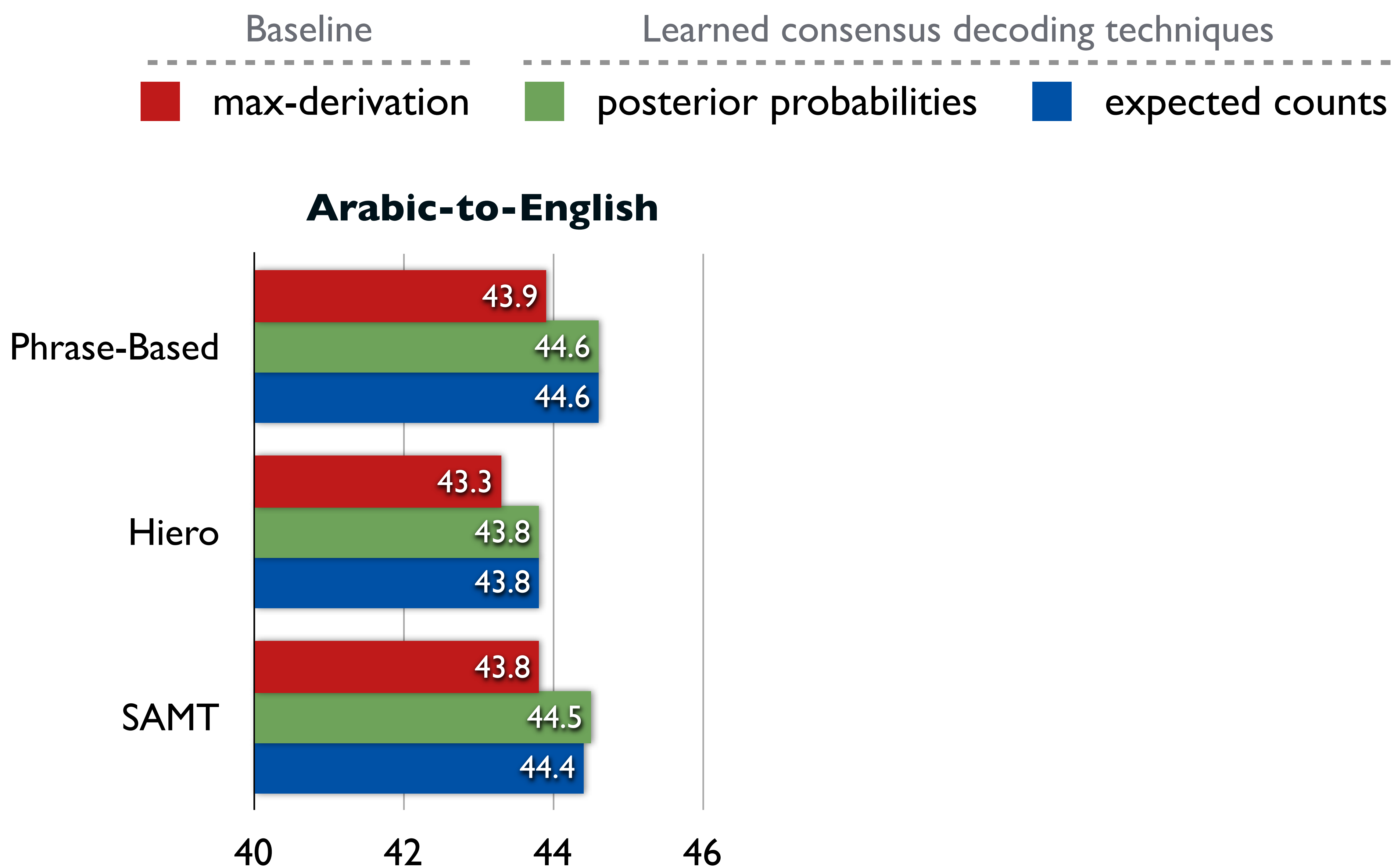
Results for Learned Consensus Decoding

Constrained data track of the 2008 NIST MT task

Baseline	Learned consensus decoding techniques	
max-derivation	posterior probabilities	expected counts

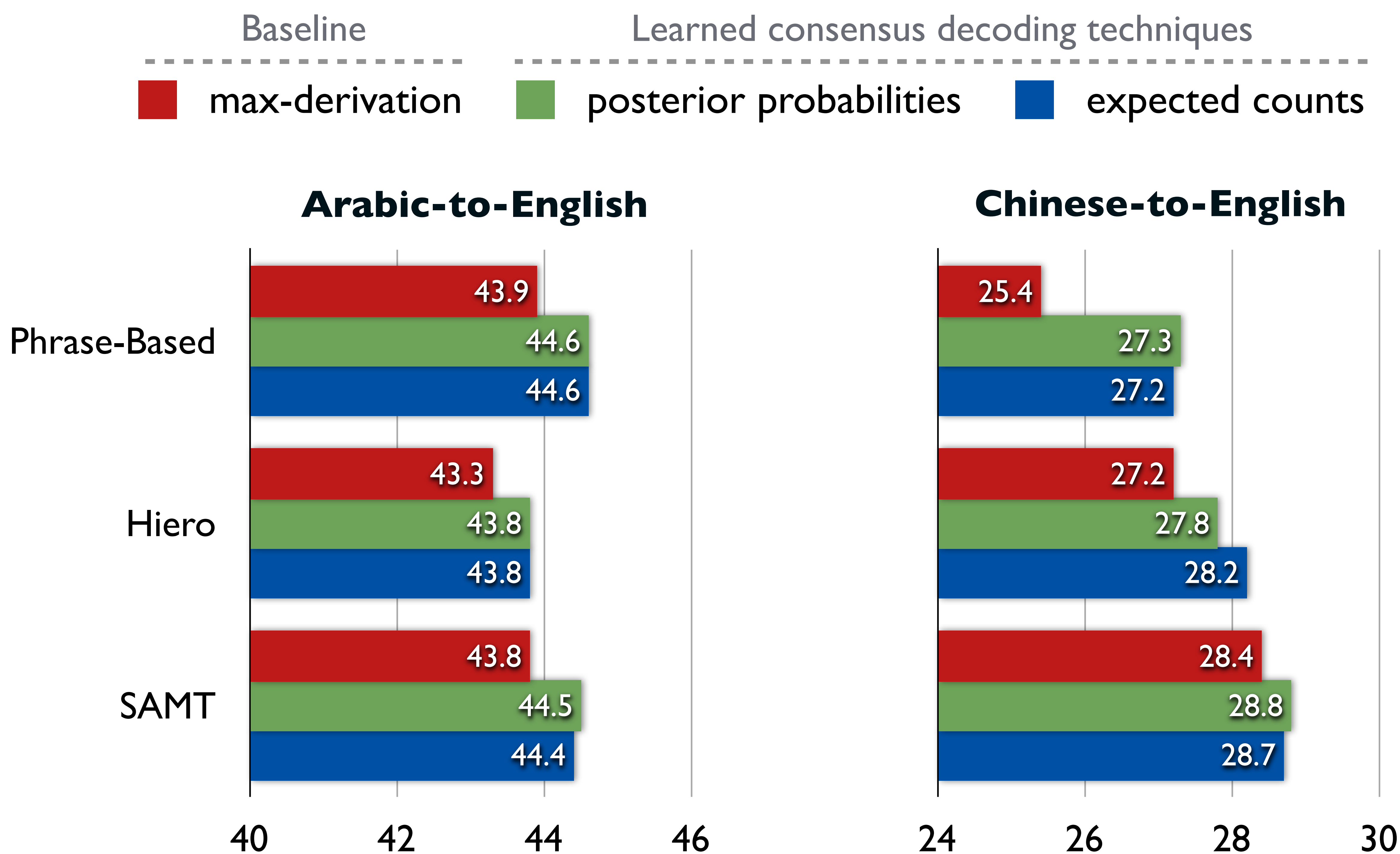
Results for Learned Consensus Decoding

Constrained data track of the 2008 NIST MT task



Results for Learned Consensus Decoding

Constrained data track of the 2008 NIST MT task



Outline

Consensus decoding review

Our model combination technique

Comparison to system combination

Outline

Consensus decoding review

Our model combination technique

All in One Slide!

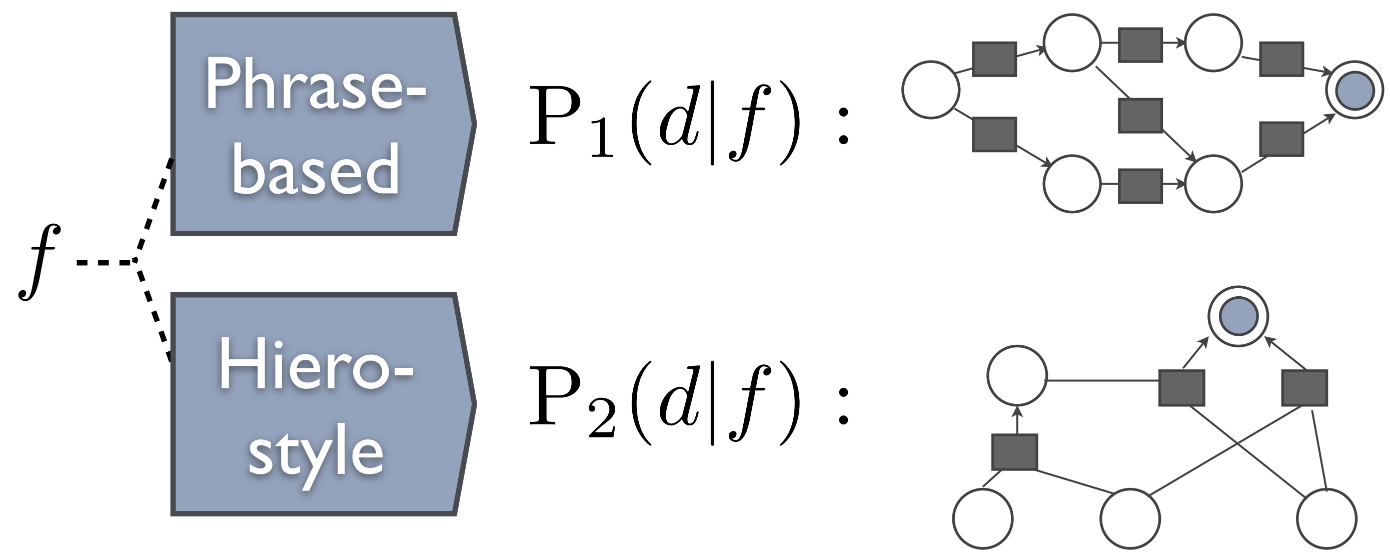
Comparison to system combination

Extending Consensus Decoding to Multiple Models

1. Build posterior forests

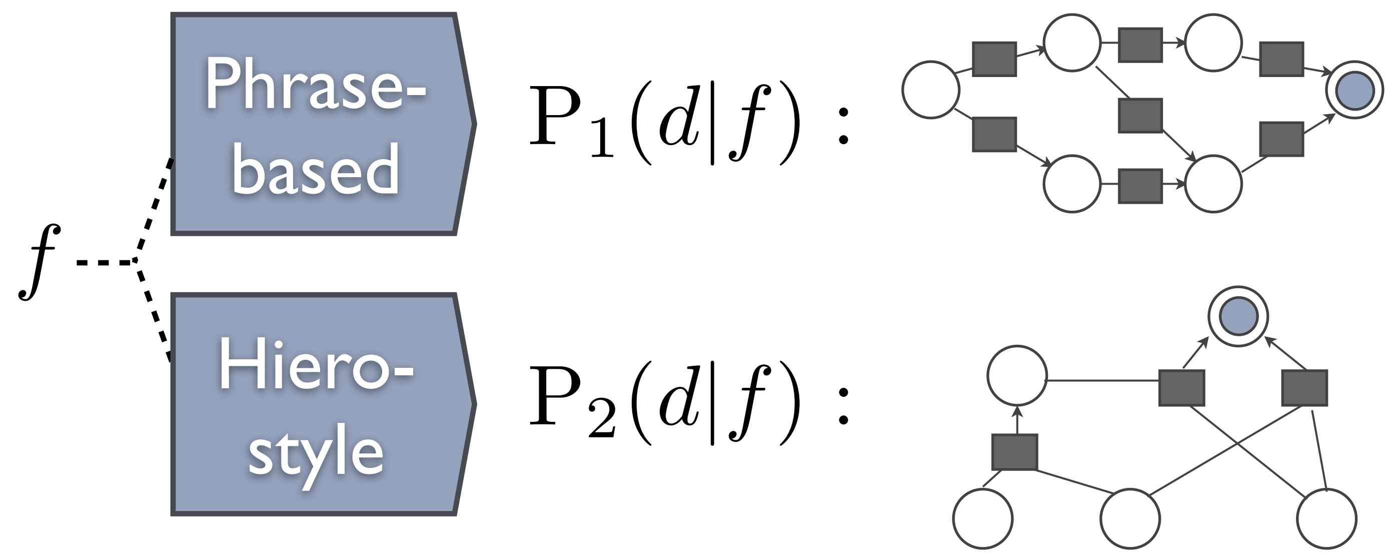
Extending Consensus Decoding to Multiple Models

1. Build posterior forests



Extending Consensus Decoding to Multiple Models

1. Build posterior forests



N-gram statistics | Length | Base model

$$\mathrm{C}(d) = \sum_{n=1}^{4} w^{(n)} \sum_{g \in n\text{-grams}} c(g,d) \cdot \mathrm{P}(g|f) + w^{(\ell)} |\sigma_e(d)| + w^{(b)} \theta \cdot \phi(d)$$

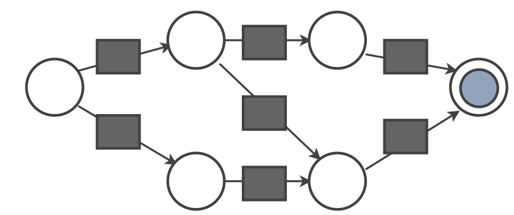
Extending Consensus Decoding to Multiple Models

1. Build posterior forests

f

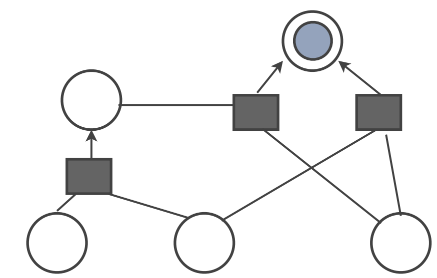
Phrase-based

$\mathrm{P}_1(d|f)$:



Hiero-style

$\mathrm{P}_2(d|f)$:



N-gram statistics

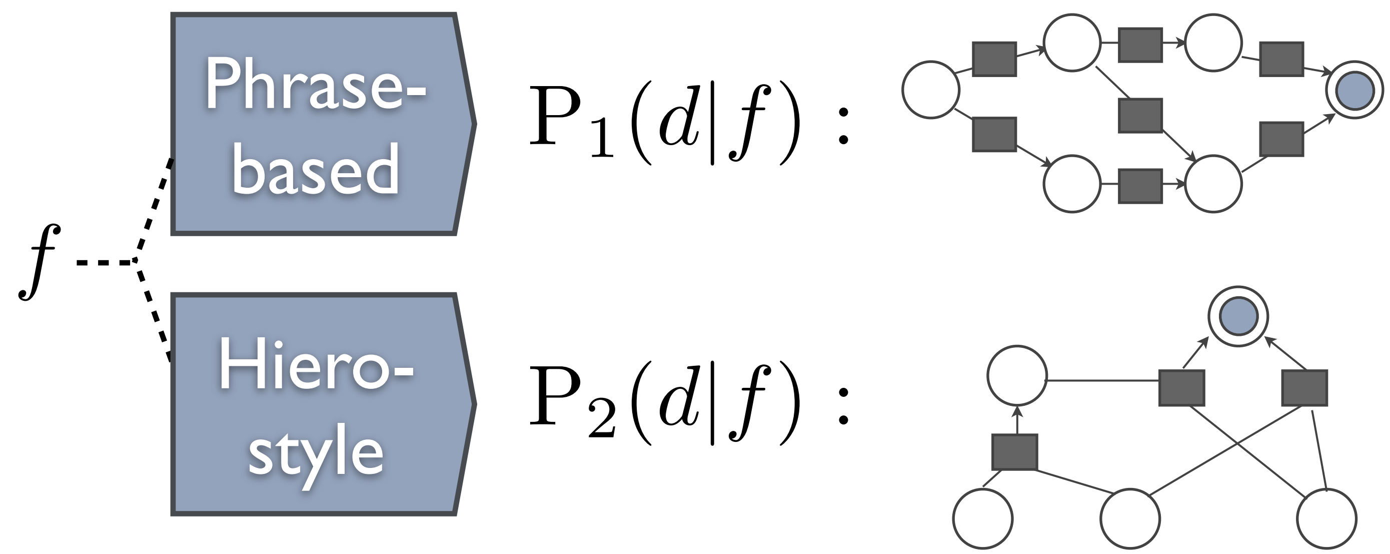
Length

Base model

$$\mathrm{C}(d) = \sum_{n=1}^{4} w^{(n)} v^{(n)}(d) + w^{(\ell)} |\sigma_e(d)| + w^{(b)} b(d)$$

Extending Consensus Decoding to Multiple Models

1. Build posterior forests

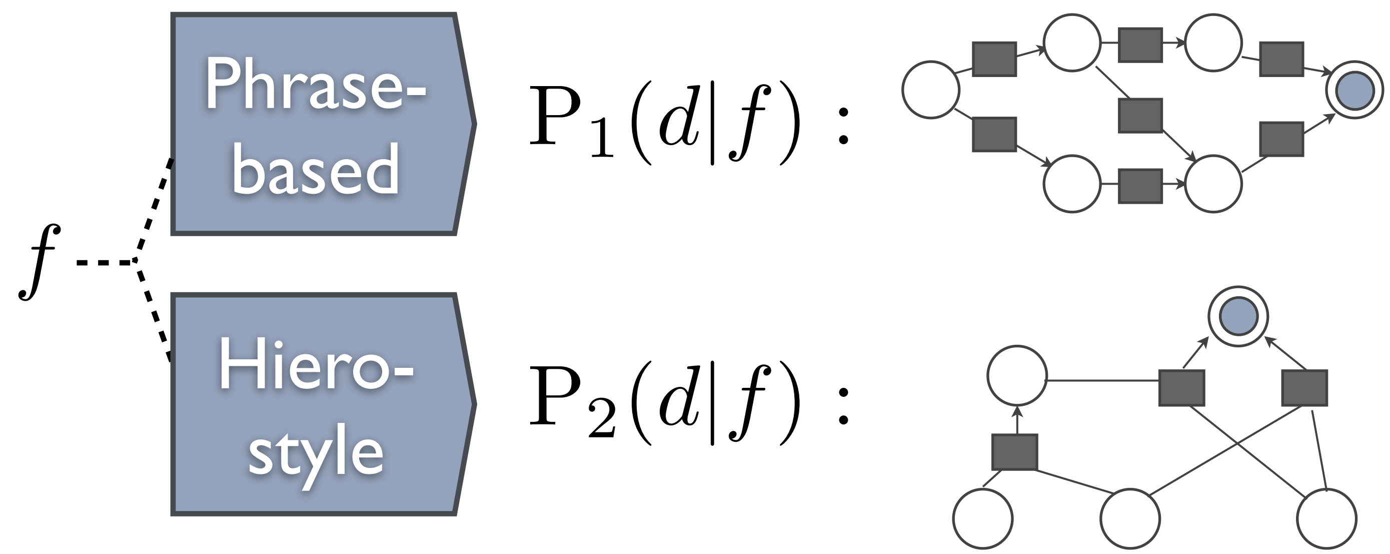


Sum over models — N-gram statistics — Length — Base model

$$\mathrm{C}(d) = \sum_{i=1}^{I} \overbrace{\sum_{n=1}^{4} w_i^{(n)} v_i^{(n)}(d)}^{\text{N-gram statistics}} + \overbrace{w^{(\ell)} |\sigma_e(d)|}^{\text{Length}} + \overbrace{w^{(b)} b(d)}^{\text{Base model}}$$

Extending Consensus Decoding to Multiple Models

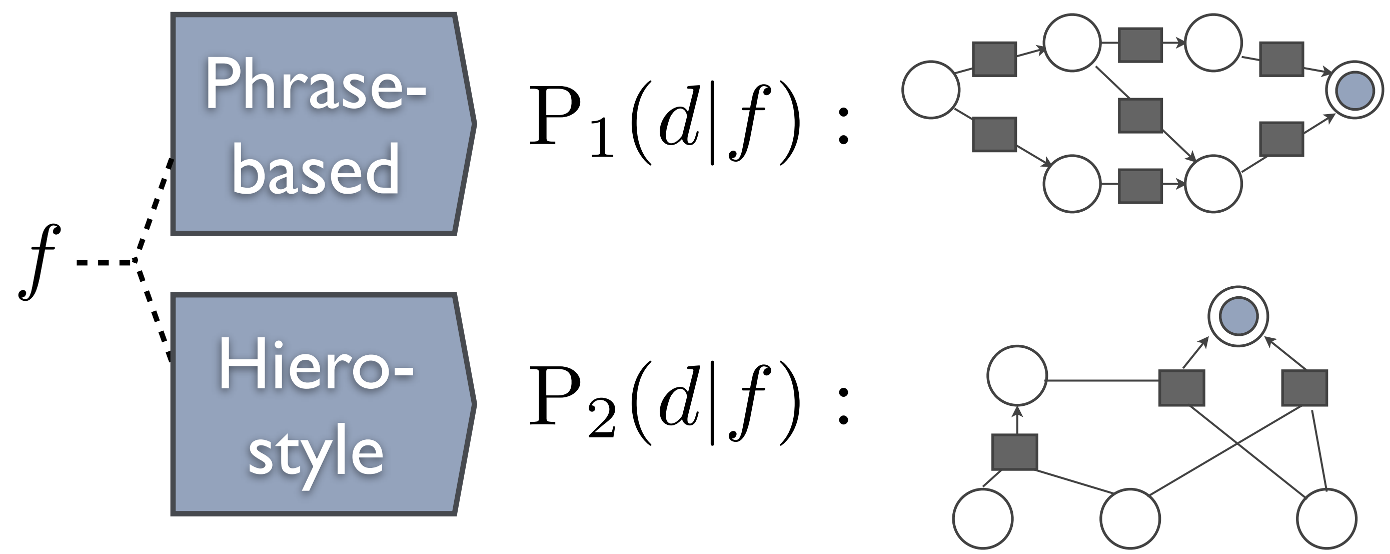
1. Build posterior forests
2. Compute n-gram statistics from forests



$$\mathrm{C}(d) = \overset{\text{Sum over models}}{\sum_{i=1}^{I}} \overbrace{\sum_{n=1}^{4} w_i^{(n)} v_i^{(n)}(d)}^{\text{N-gram statistics}} \overbrace{+ w^{(\ell)} |\sigma_e(d)|}^{\text{Length}} \overbrace{+ w^{(b)} b(d)}^{\text{Base model}}$$

Extending Consensus Decoding to Multiple Models

1. Build posterior forests
2. Compute n-gram statistics from forests
3. Union all forests

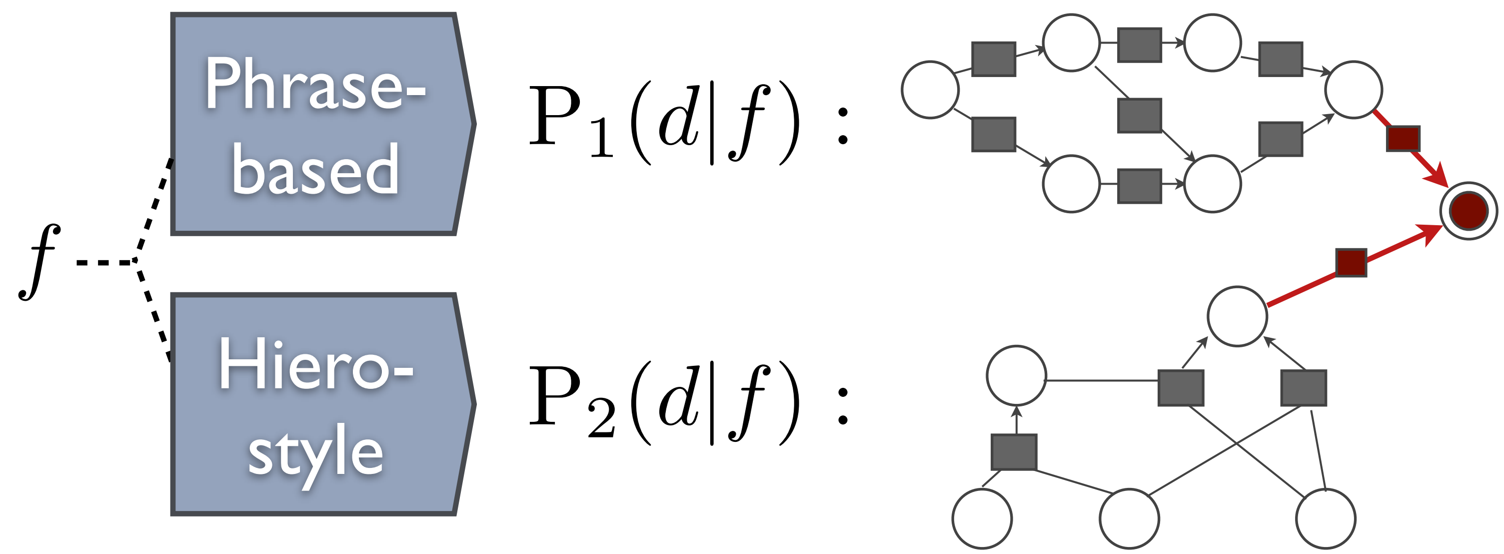


Sum over models — N-gram statistics — Length — Base model

$$\mathrm{C}(d) = \sum_{i=1}^{I} \sum_{n=1}^{4} w_i^{(n)} v_i^{(n)}(d) + w^{(\ell)} |\sigma_e(d)| + w^{(b)} b(d)$$

Extending Consensus Decoding to Multiple Models

1. Build posterior forests
2. Compute n-gram statistics from forests
3. Union all forests

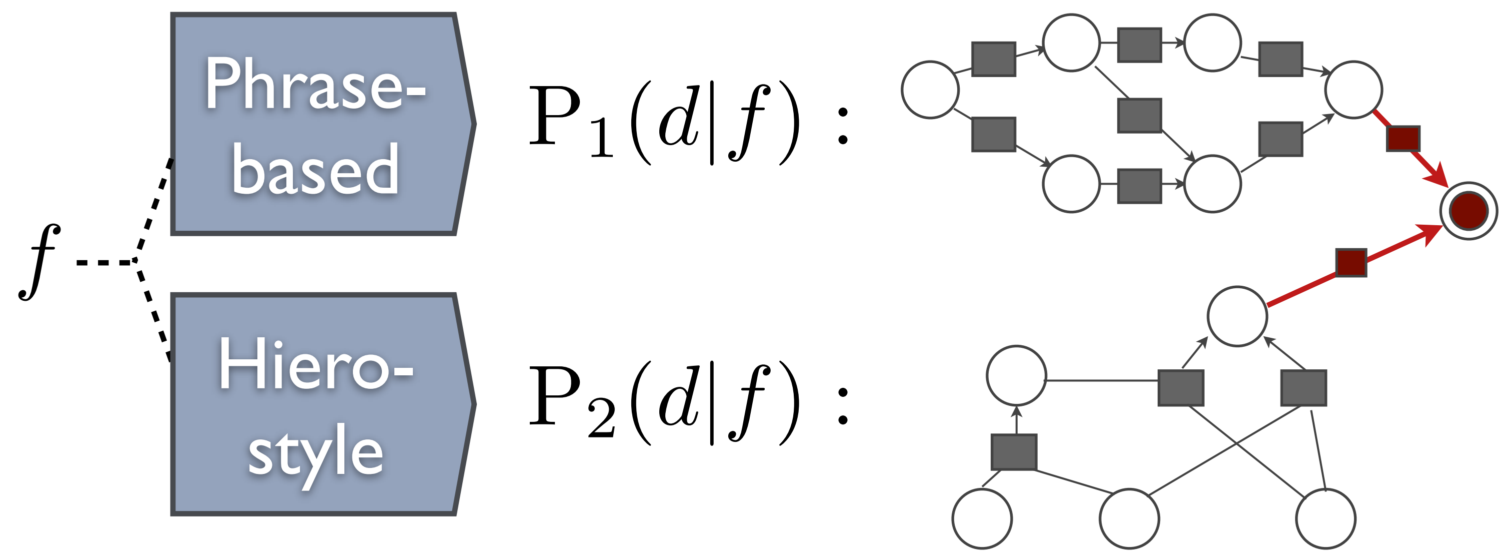


Sum over models — N-gram statistics — Length — Base model

$$\mathrm{C}(d) = \sum_{i=1}^{I} \sum_{n=1}^{4} w_i^{(n)} v_i^{(n)}(d) + w^{(\ell)} |\sigma_e(d)| + w^{(b)} b(d)$$

Extending Consensus Decoding to Multiple Models

1. Build posterior forests
2. Compute n-gram statistics from forests
3. Union all forests

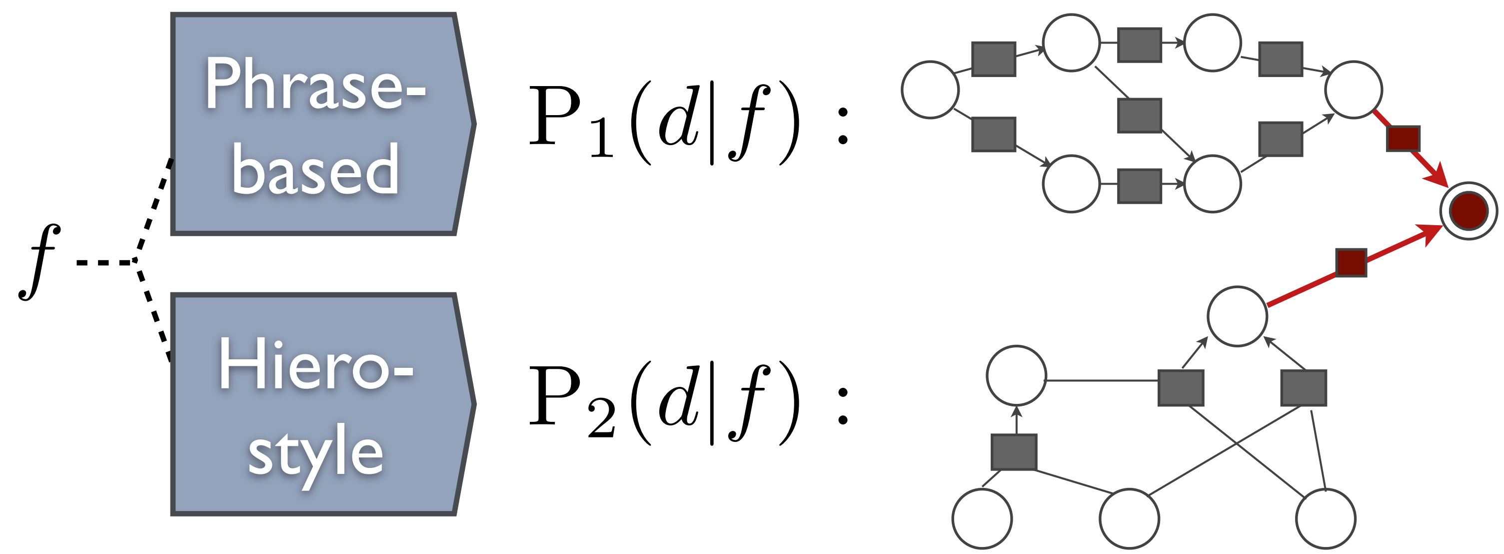


Sum over models | N-gram statistics | Model choice | Length | Base model

$$C(d) = \sum_{i=1}^{I} \sum_{n=1}^{4} w_i^{(n)} v_i^{(n)}(d) + w_i^{(\alpha)} \alpha_i(d) + w^{(\ell)} |\sigma_e(d)| + w^{(b)} b(d)$$

Extending Consensus Decoding to Multiple Models

1. Build posterior forests
2. Compute n-gram statistics from forests
3. Union all forests



Sum over models | N-gram statistics | Model choice | Length | Base model

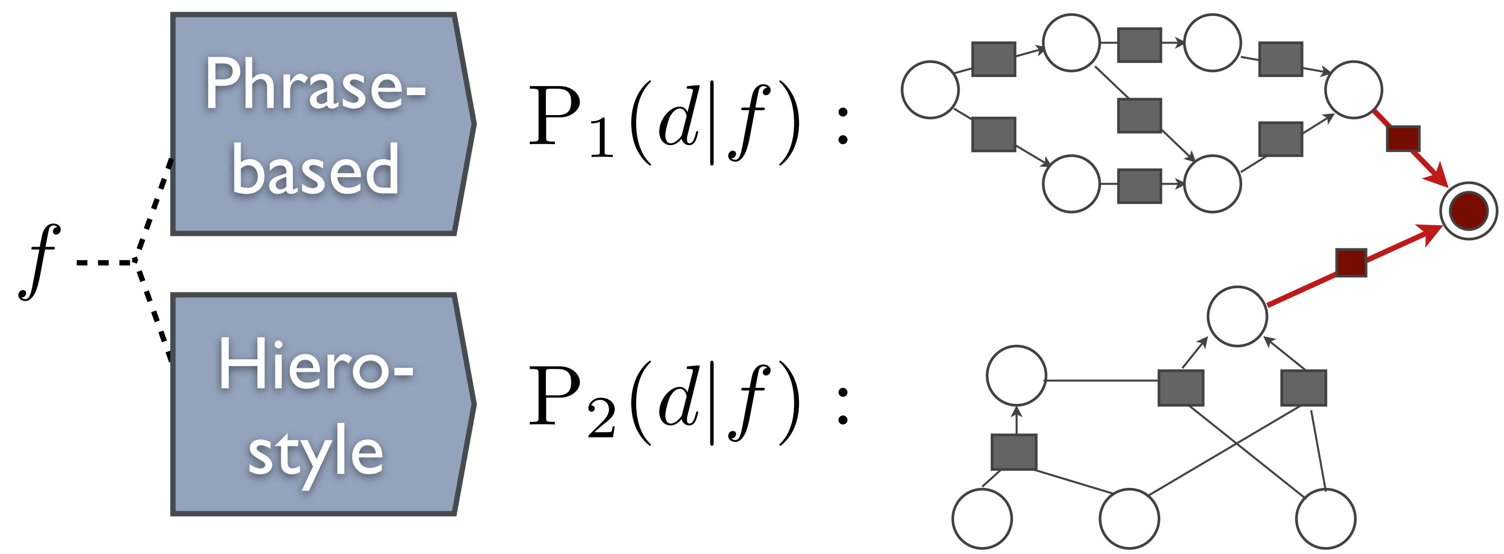
$$C(d) = \sum_{i=1}^{I} \underbrace{\sum_{n=1}^{4} w_i^{(n)} v_i^{(n)}(d)}_{\text{N-gram statistics}} + \underbrace{w_i^{(\alpha)} \alpha_i(d)}_{\text{Model choice}} + \underbrace{w^{(\ell)} |\sigma_e(d)|}_{\text{Length}} + \underbrace{w^{(b)} b(d)}_{\text{Base model}}$$

Model choice: *Indicator feature* for the system that originally generated d

Base model: Model score under the system that generated d

Extending Consensus Decoding to Multiple Models

1. Build posterior forests
2. Compute n-gram statistics from forests
3. Union all forests
4. Optimize multi-model consensus objective



Sum over models | N-gram statistics | Model choice | Length | Base model

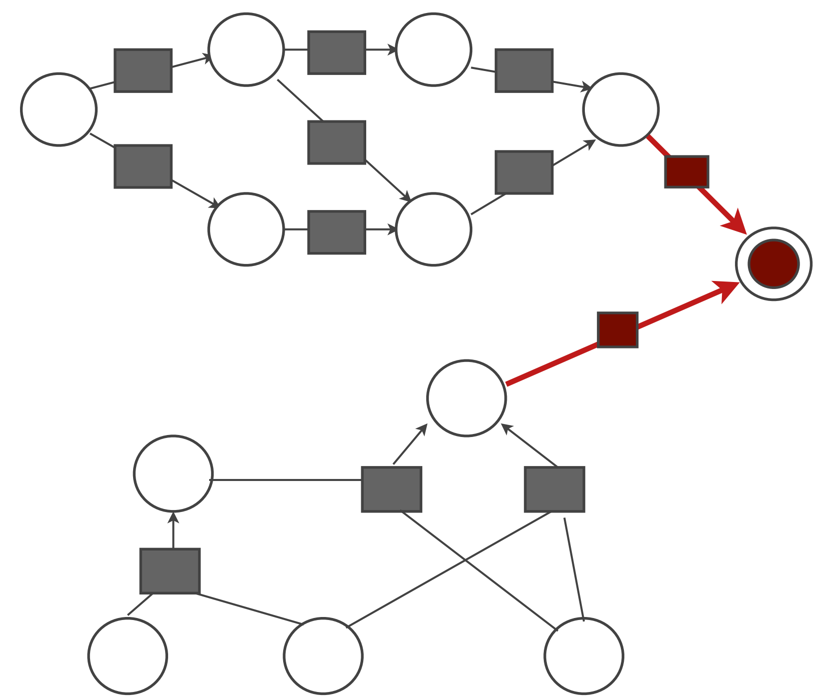
$$C(d) = \sum_{i=1}^{I} \sum_{n=1}^{4} w_i^{(n)} v_i^{(n)}(d) + w_i^{(\alpha)} \alpha_i(d) + w^{(\ell)} |\sigma_e(d)| + w^{(b)} b(d)$$

Model choice: *Indicator feature* for the system that originally generated d

Base model: Model score under the system that generated d

Extending Consensus Decoding to Multiple Models

1. Build posterior forests
2. Compute n-gram statistics from forests
3. Union all forests
4. Optimize multi-model consensus objective



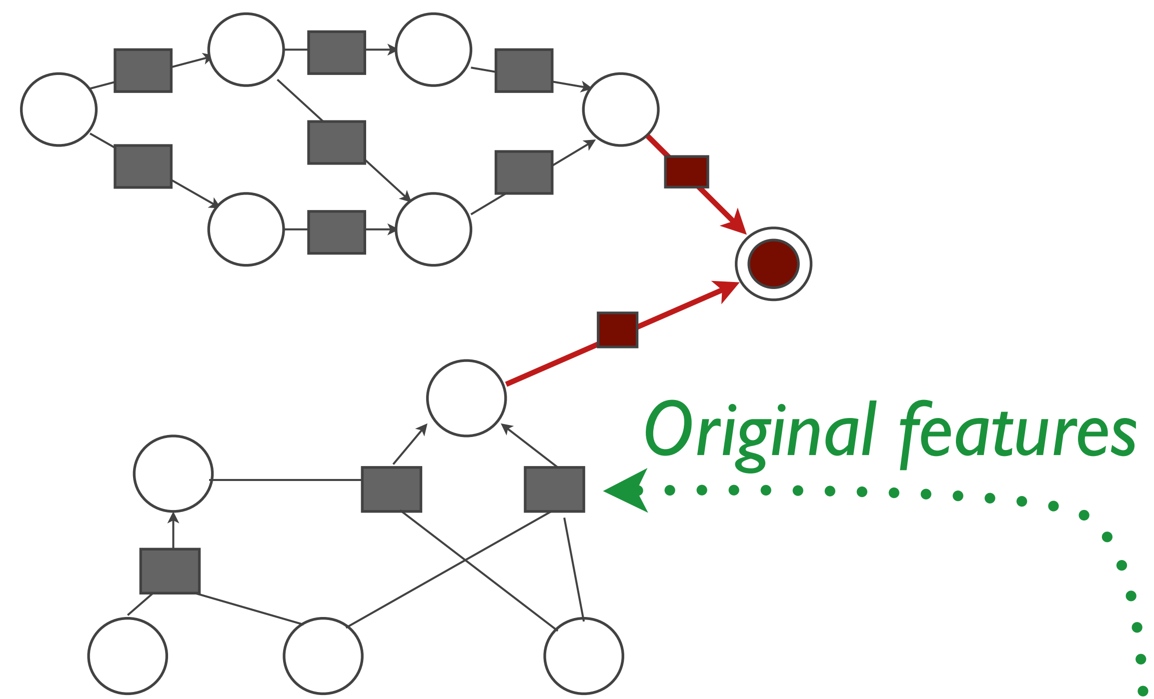
$$\mathrm{C}(d) = \underbrace{\sum_{i=1}^{I}}_{\text{Sum over models}} \underbrace{\sum_{n=1}^{4} w_i^{(n)} v_i^{(n)}(d)}_{\text{N-gram statistics}} + \underbrace{w_i^{(\alpha)} \alpha_i(d)}_{\text{Model choice}} + \underbrace{w^{(\ell)} |\sigma_e(d)|}_{\text{Length}} + \underbrace{w^{(b)} b(d)}_{\text{Base model}}$$

Model choice: *Indicator feature* for the system that originally generated d

Base model: Model score under the system that generated d

Extending Consensus Decoding to Multiple Models

1. Build posterior forests
2. Compute n-gram statistics from forests
3. Union all forests
4. **Optimize multi-model consensus objective**



Sum over models | N-gram statistics | Model choice | Length | Base model

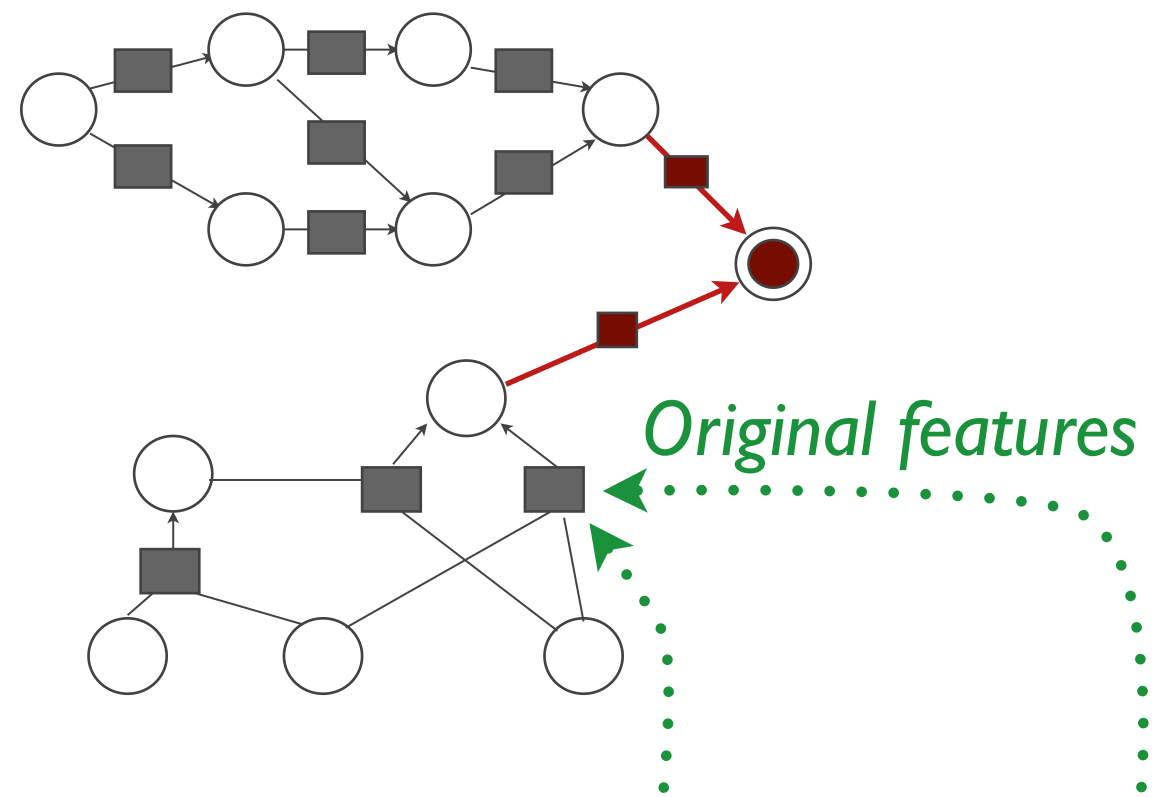
$$C(d) = \sum_{i=1}^{I} \sum_{n=1}^{4} w_i^{(n)} v_i^{(n)}(d) + w_i^{(\alpha)} \alpha_i(d) + w^{(\ell)} |\sigma_e(d)| + w^{(b)} b(d)$$

Model choice: *Indicator feature* for the system that originally generated d

Base model: Model score under the system that generated d

Extending Consensus Decoding to Multiple Models

1. Build posterior forests
2. Compute n-gram statistics from forests
3. Union all forests
4. **Optimize multi-model consensus objective**



Sum over models | N-gram statistics | Model choice | Length | Base model

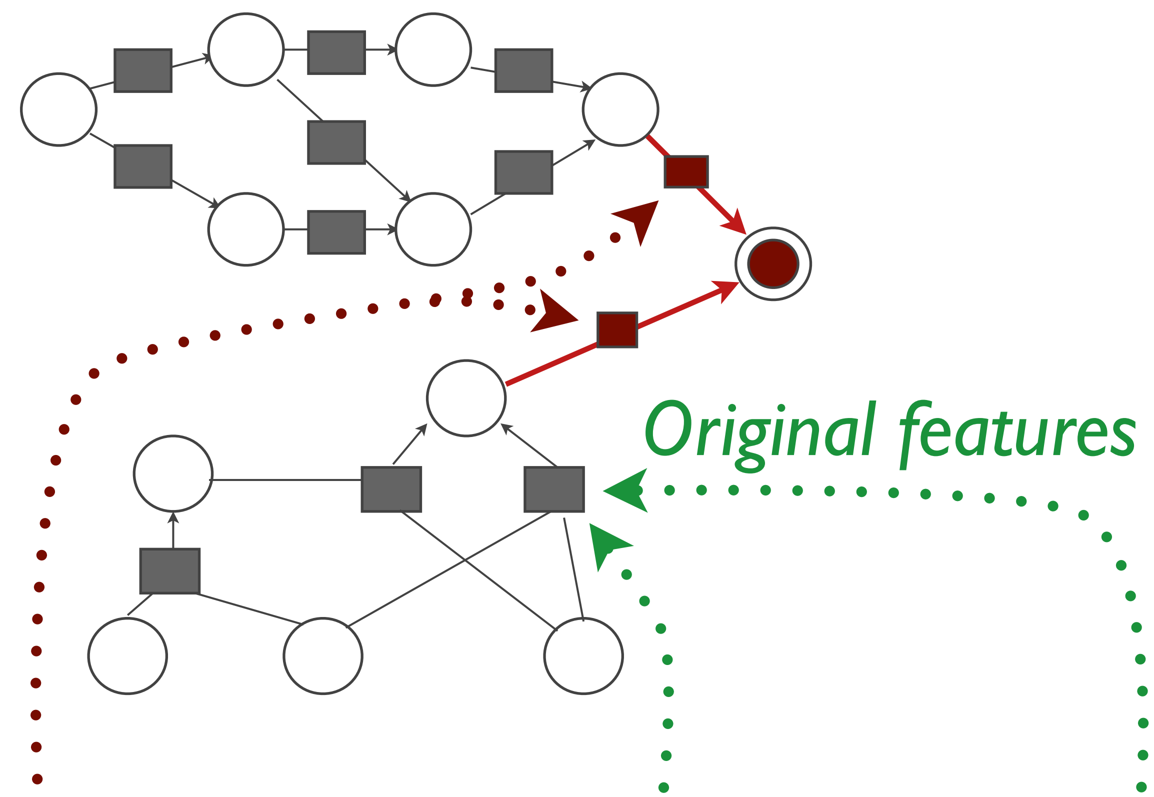
$$C(d) = \sum_{i=1}^{I} \sum_{n=1}^{4} w_i^{(n)} v_i^{(n)}(d) + w_i^{(\alpha)} \alpha_i(d) + w^{(\ell)} |\sigma_e(d)| + w^{(b)} b(d)$$

Model choice: *Indicator feature* for the system that originally generated d

Base model: Model score under the system that generated d

Extending Consensus Decoding to Multiple Models

1. Build posterior forests
2. Compute n-gram statistics from forests
3. Union all forests
4. Optimize multi-model consensus objective



Sum over models — N-gram statistics — Model choice — Length — Base model

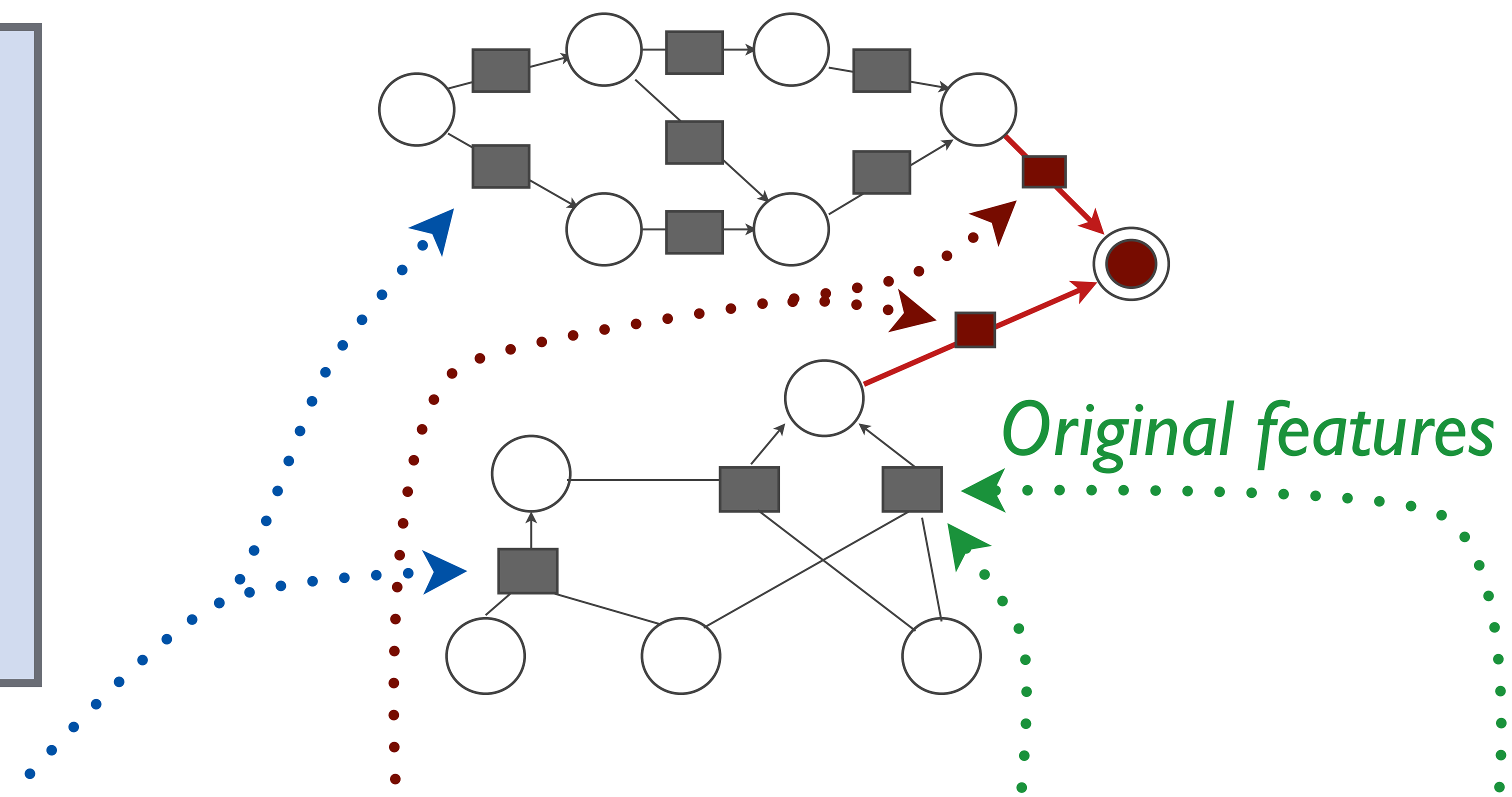
$$C(d) = \sum_{i=1}^{I} \sum_{n=1}^{4} w_i^{(n)} v_i^{(n)}(d) + w_i^{(\alpha)} \alpha_i(d) + w^{(\ell)} |\sigma_e(d)| + w^{(b)} b(d)$$

Model choice: *Indicator feature* for the system that originally generated d

Base model: Model score under the system that generated d

Extending Consensus Decoding to Multiple Models

1. Build posterior forests
2. Compute n-gram statistics from forests
3. Union all forests
4. Optimize multi-model consensus objective



Sum over models | N-gram statistics | Model choice | Length | Base model

$$\mathrm{C}(d) = \sum_{i=1}^{I} \sum_{n=1}^{4} w_i^{(n)} v_i^{(n)}(d) + w_i^{(\alpha)} \alpha_i(d) + w^{(\ell)} |\sigma_e(d)| + w^{(b)} b(d)$$

Model choice: *Indicator feature* for the system that originally generated d

Base model: Model score under the system that generated d

Properties of Model Combination

Sum over models | N-gram statistics | Model choice | Length | Base model

$$\mathrm{C}(d) = \sum_{i=1}^{I} \sum_{n=1}^{4} w_i^{(n)} v_i^{(n)}(d) + w_i^{(\alpha)} \alpha_i(d) + w^{(\ell)} |\sigma_e(d)| + w^{(b)} b(d)$$

Properties of Model Combination

Sum over models — N-gram statistics — Model choice — Length — Base model

$$\mathrm{C}(d) = \sum_{i=1}^{I} \sum_{n=1}^{4} w_i^{(n)} v_i^{(n)}(d) + w_i^{(\alpha)} \alpha_i(d) + w^{(\ell)} |\sigma_e(d)| + w^{(b)} b(d)$$

- Reduces to consensus decoding when we have only one model

Properties of Model Combination

Sum over models — N-gram statistics — Model choice — Length — Base model

$$\mathrm{C}(d) = \sum_{i=1}^{I} \overbrace{\sum_{n=1}^{4} w_i^{(n)} v_i^{(n)}(d)}^{\text{N-gram statistics}} + \overbrace{w_i^{(\alpha)} \alpha_i(d)}^{\text{Model choice}} + \overbrace{w^{(\ell)} |\sigma_e(d)|}^{\text{Length}} + \overbrace{w^{(b)} b(d)}^{\text{Base model}}$$

- Reduces to consensus decoding when we have only one model
- A linear model: w can be tuned to maximize output performance

Properties of Model Combination

Sum over models N-gram statistics Model choice Length Base model

$$C(d) = \sum_{i=1}^{I} \sum_{n=1}^{4} w_i^{(n)} v_i^{(n)}(d) + w_i^{(\alpha)} \alpha_i(d) + w^{(\ell)} |\sigma_e(d)| + w^{(b)} b(d)$$

- Reduces to consensus decoding when we have only one model
- A linear model: w can be tuned to maximize output performance
- No concept of a primary system

Properties of Model Combination

Sum over models | N-gram statistics | Model choice | Length | Base model

$$\mathrm{C}(d) = \sum_{i=1}^{I} \underbrace{\sum_{n=1}^{4} w_i^{(n)} v_i^{(n)}(d)}_{\text{N-gram statistics}} + \underbrace{w_i^{(\alpha)} \alpha_i(d)}_{\text{Model choice}} + \underbrace{w^{(\ell)} |\sigma_e(d)|}_{\text{Length}} + \underbrace{w^{(b)} b(d)}_{\text{Base model}}$$

- Reduces to consensus decoding when we have only one model
- A linear model: w can be tuned to maximize output performance
- No concept of a primary system
- Every possible output was a derivation under *some original model*

Model Combination Experimental Results

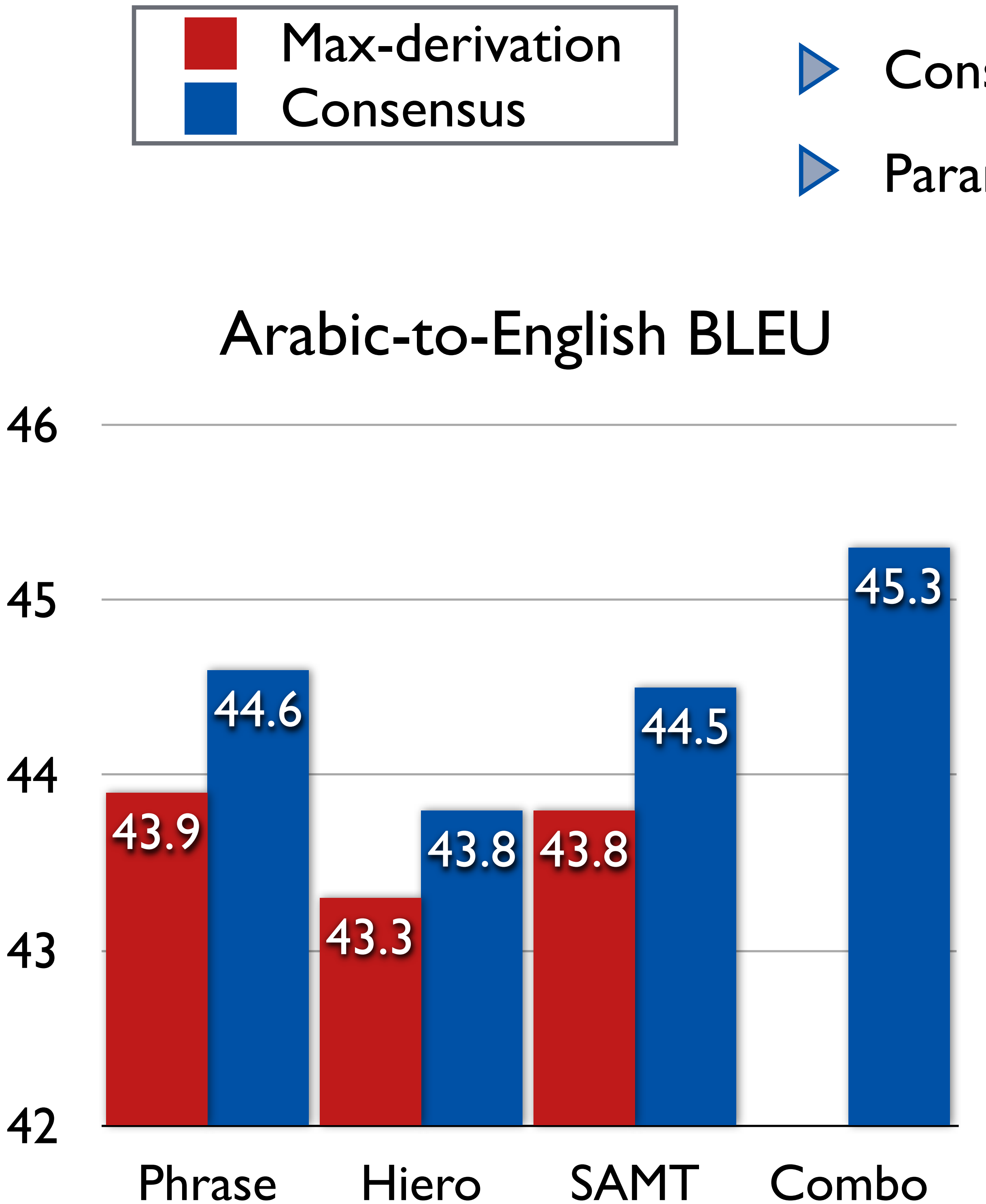
■	Max-derivation
■	Consensus

- Compared three in-house Google systems
- Constrained data track of 2008 NIST task
- Parameters tuned on NIST 2004 eval set

Model Combination Experimental Results

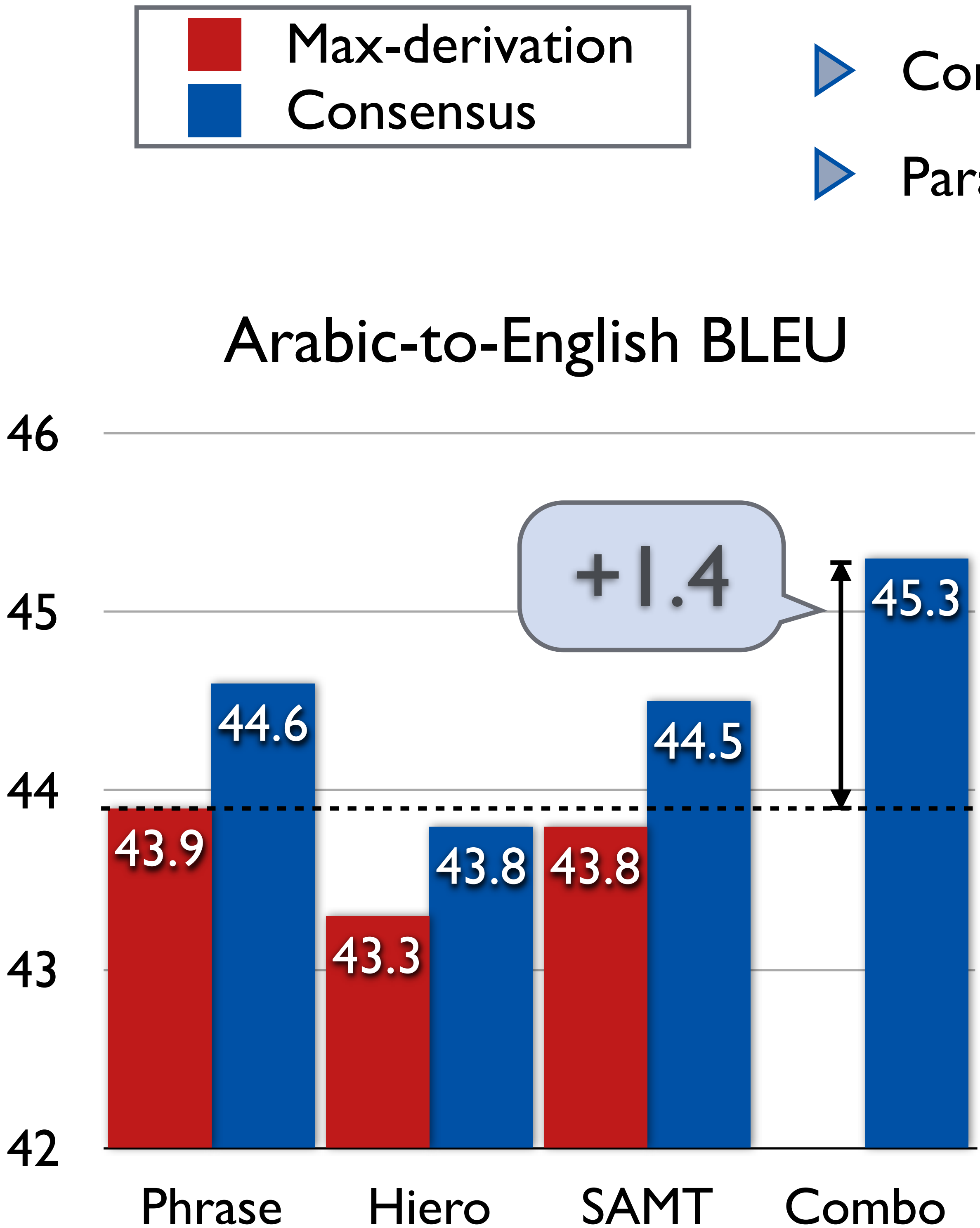
Gooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooogle

- Compared three in-house Google systems
- Constrained data track of 2008 NIST task
- Parameters tuned on NIST 2004 eval set



Model Combination Experimental Results

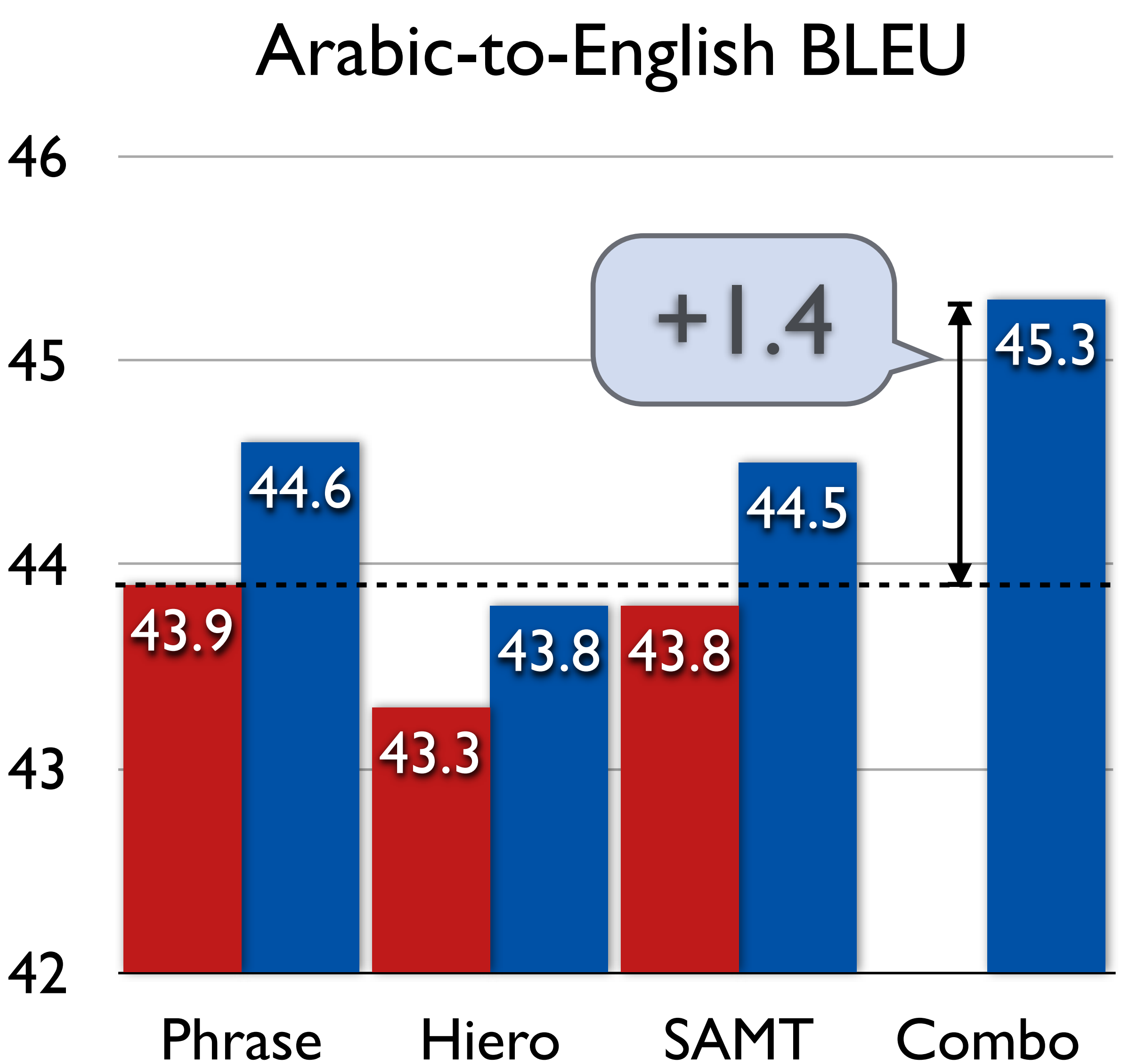
- Compared three in-house Google systems
- Constrained data track of 2008 NIST task
- Parameters tuned on NIST 2004 eval set

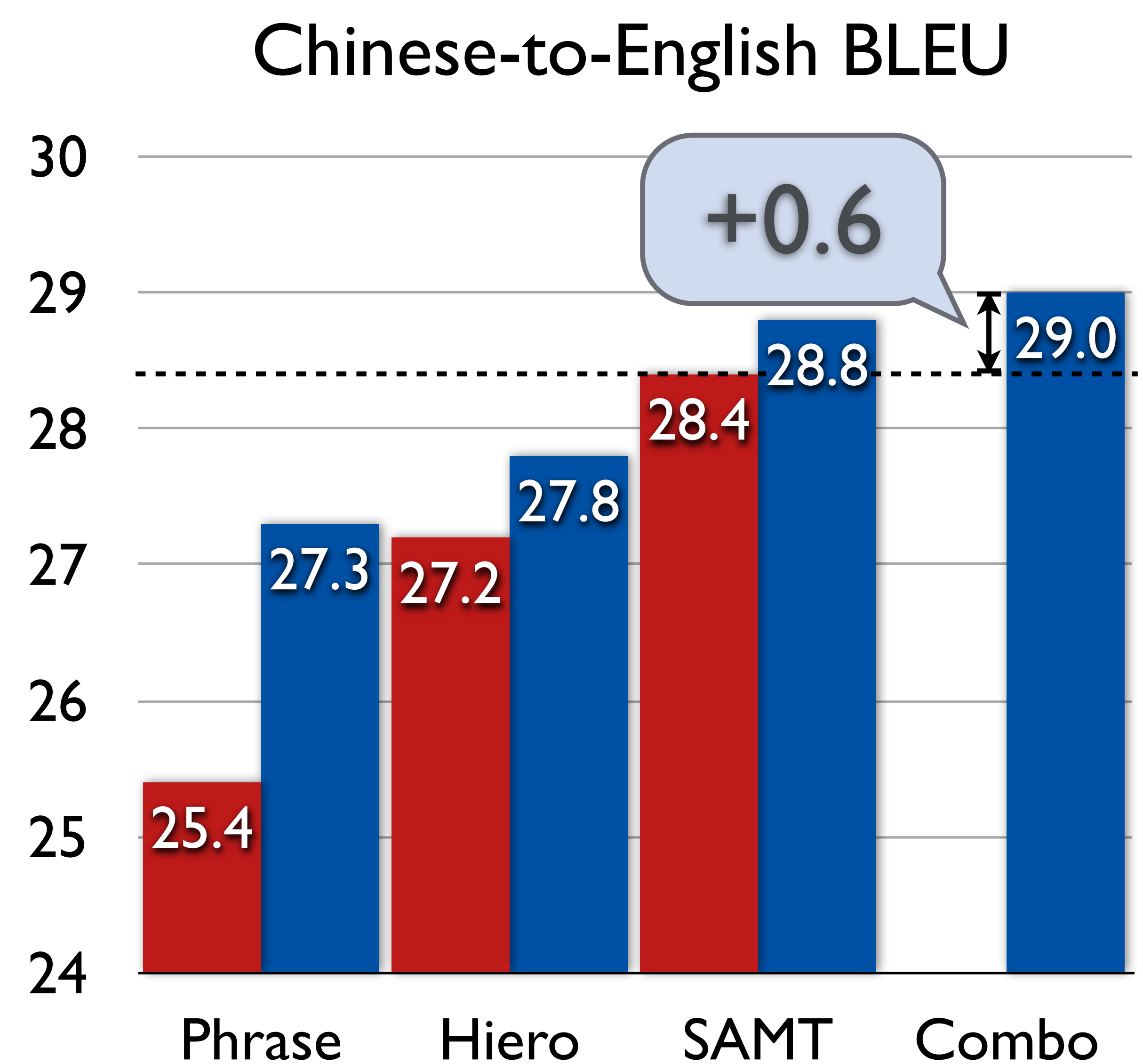


Model Combination Experimental Results

Max-derivation
Consensus

- Compared three in-house Google systems
- Constrained data track of 2008 NIST task
- Parameters tuned on NIST 2004 eval set

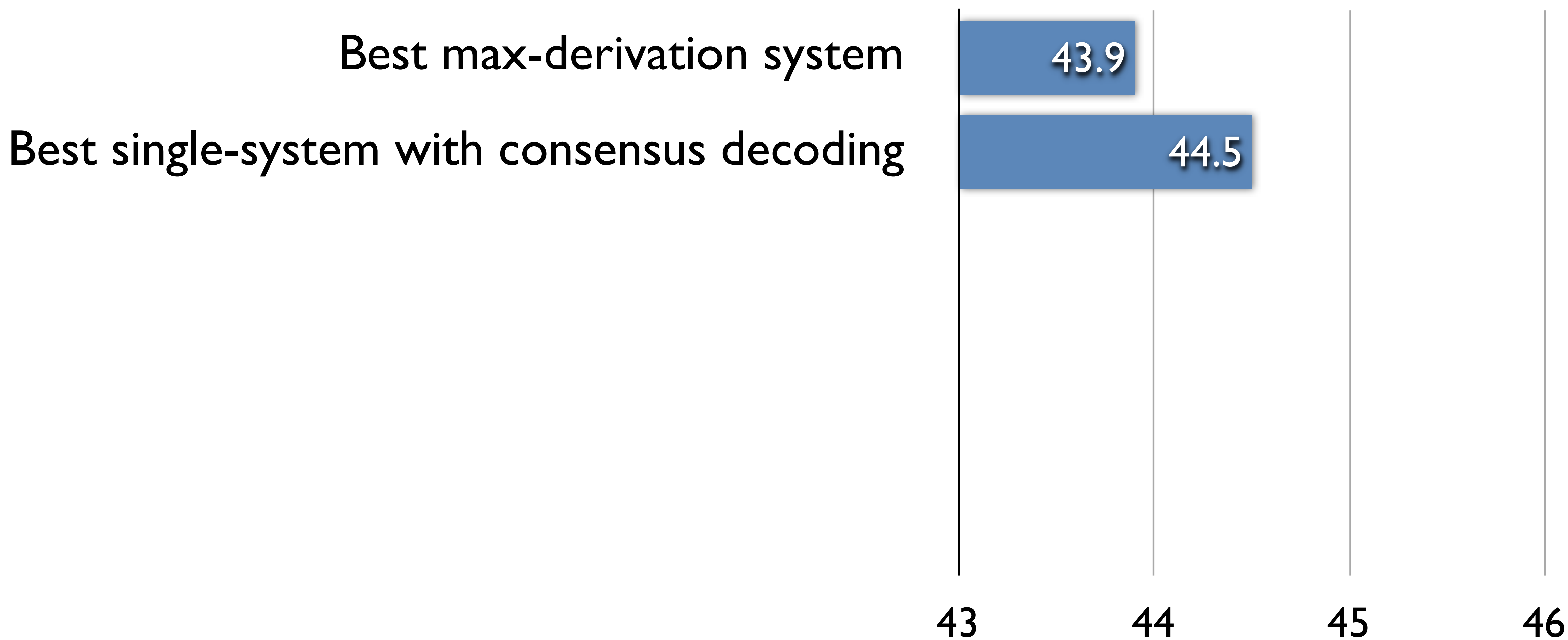




Sources of Improvement: Arabic-to-English

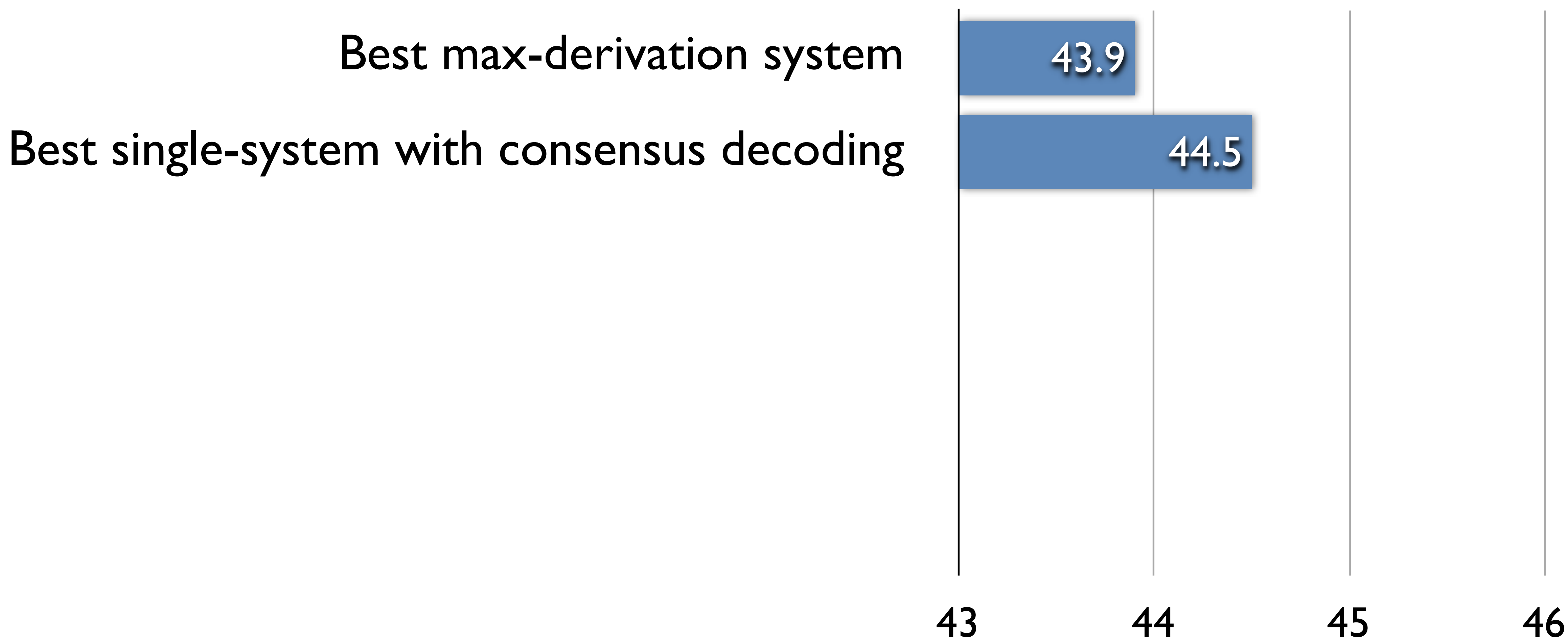
Goooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooogle

Sources of Improvement: Arabic-to-English



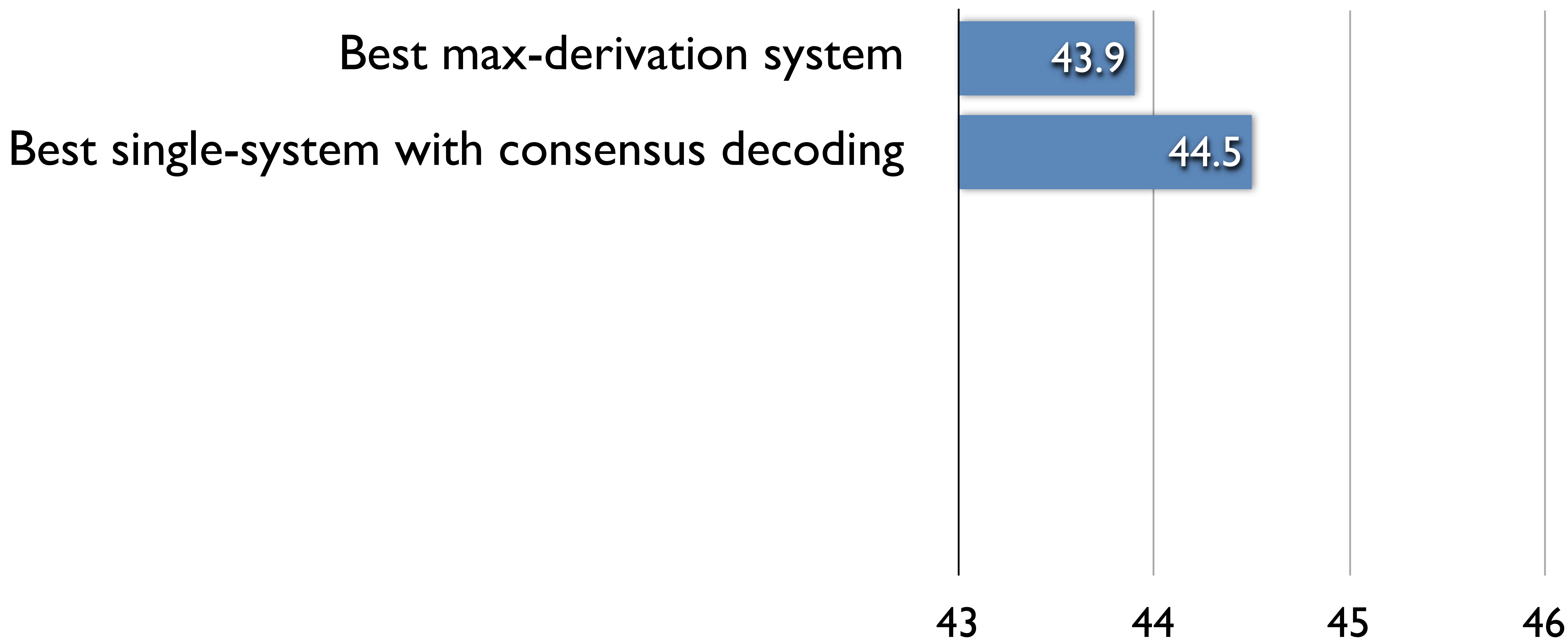
Sources of Improvement: Arabic-to-English

- Do we need model choice features?



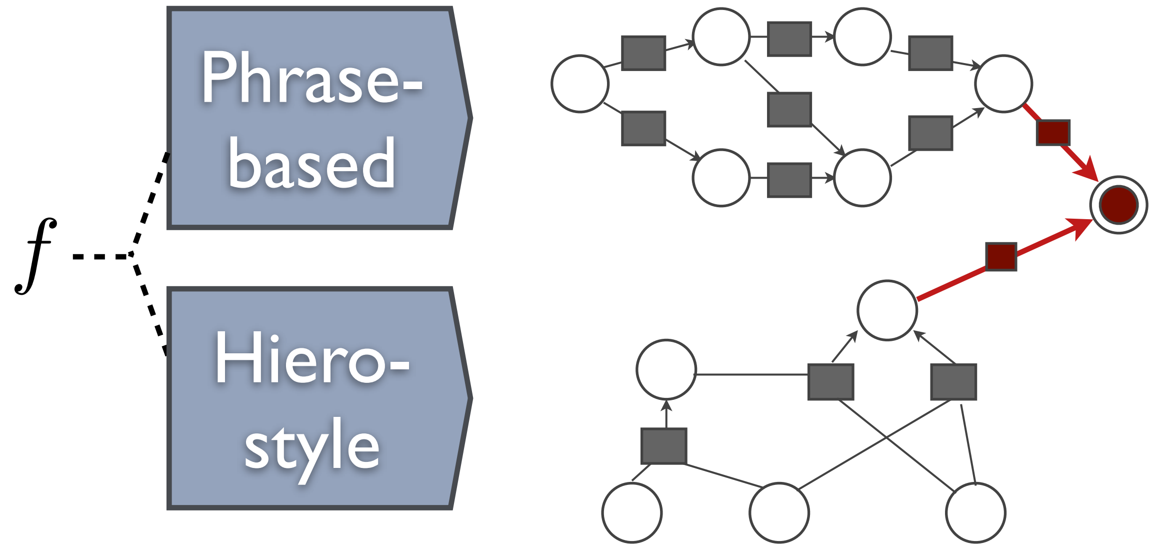
Sources of Improvement: Arabic-to-English

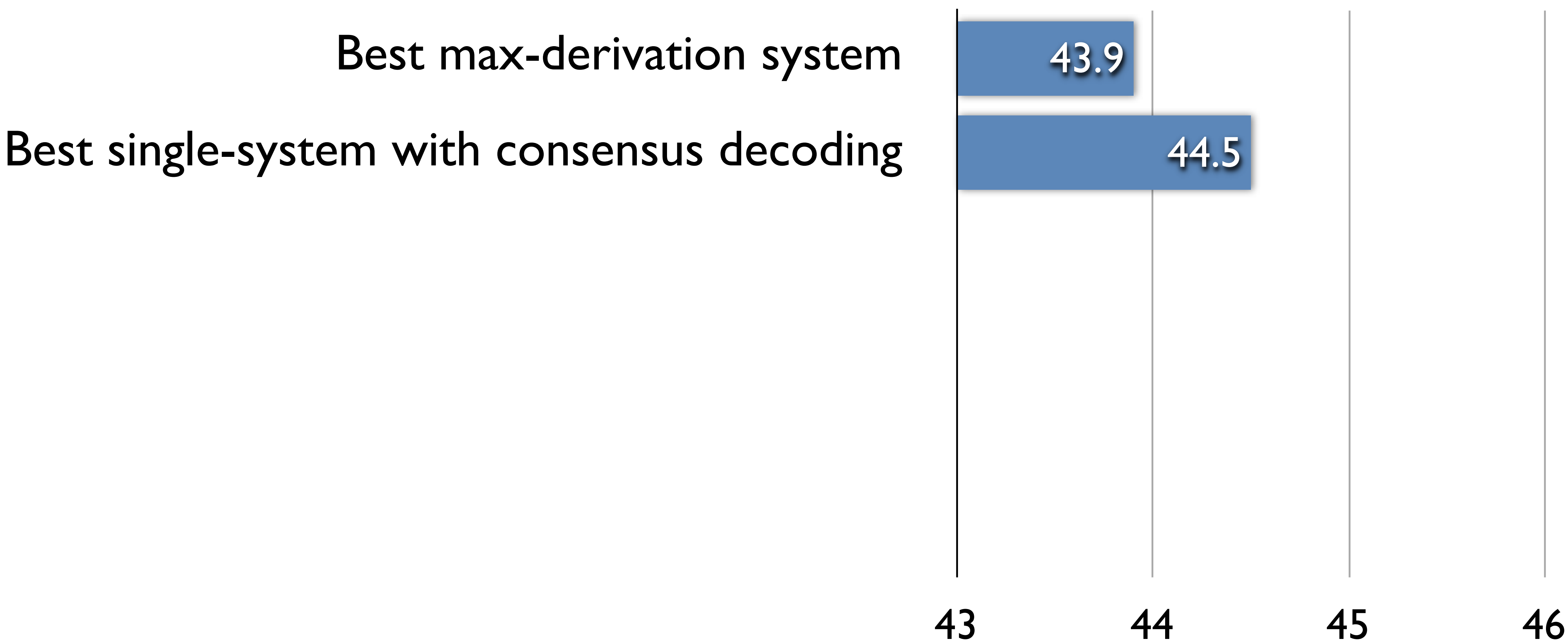
- Do we need model choice features?
- Should n-gram statistics be model specific?



Sources of Improvement: Arabic-to-English

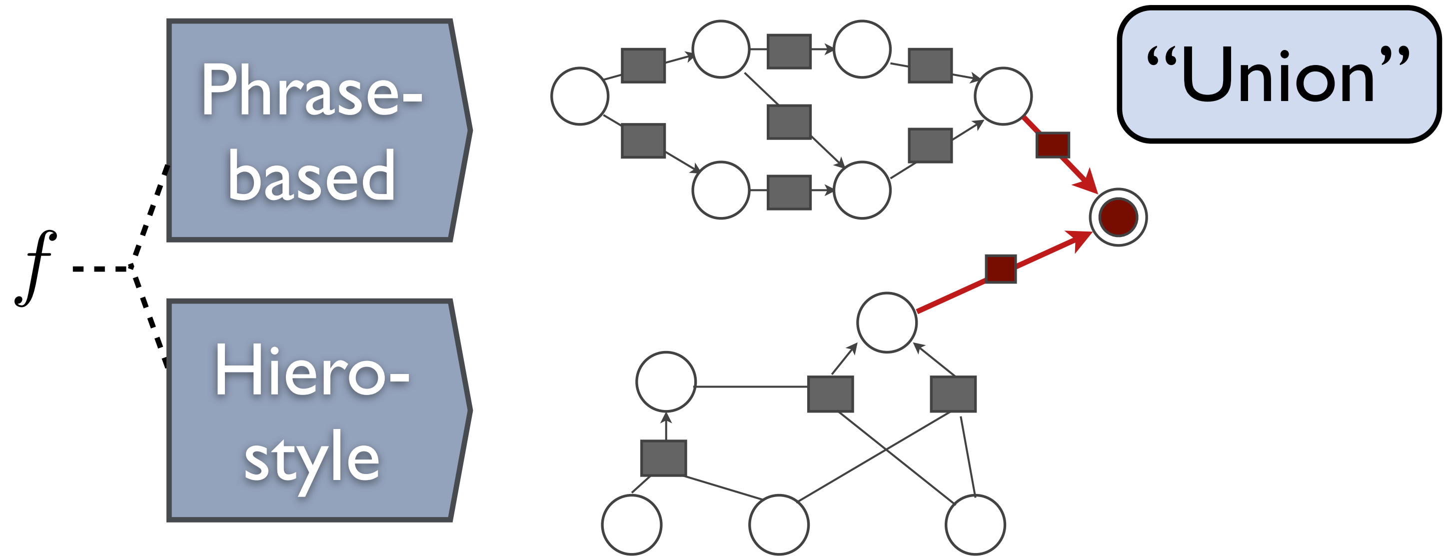
- Do we need model choice features?
- Should n-gram statistics be model specific?

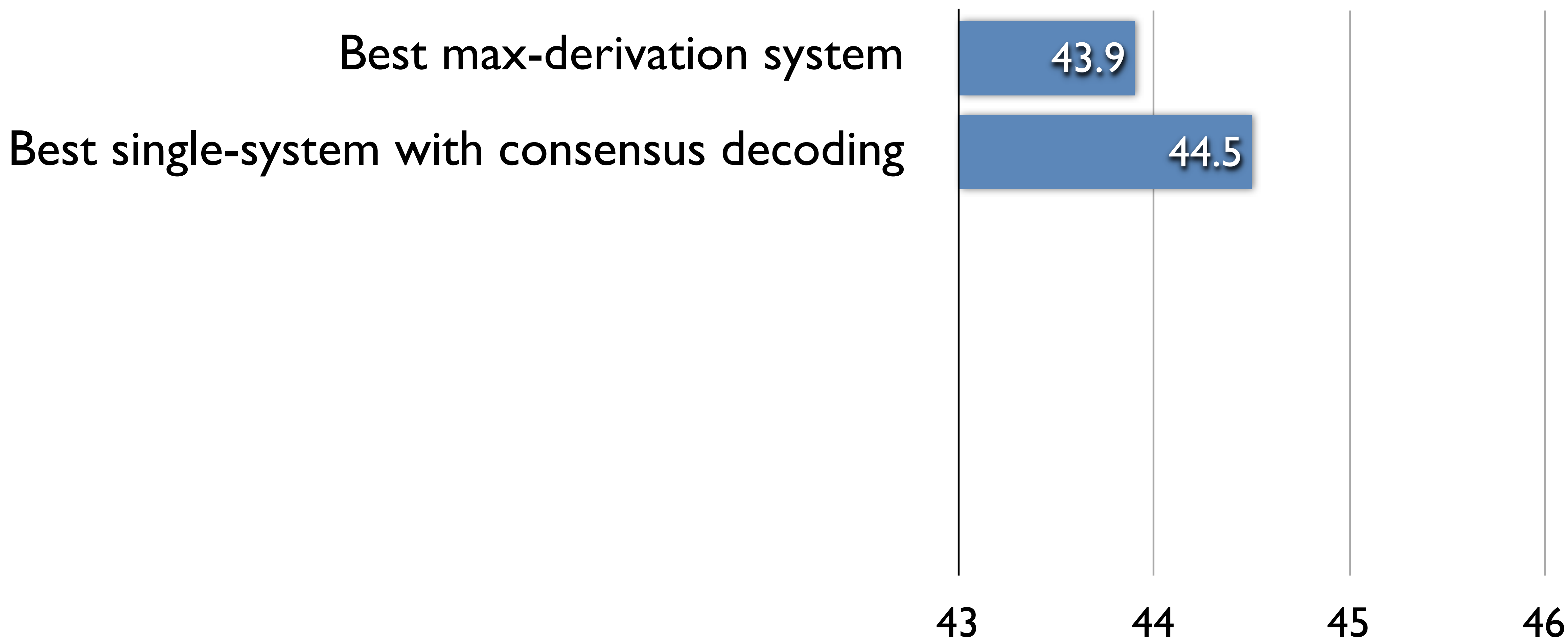




Sources of Improvement: Arabic-to-English

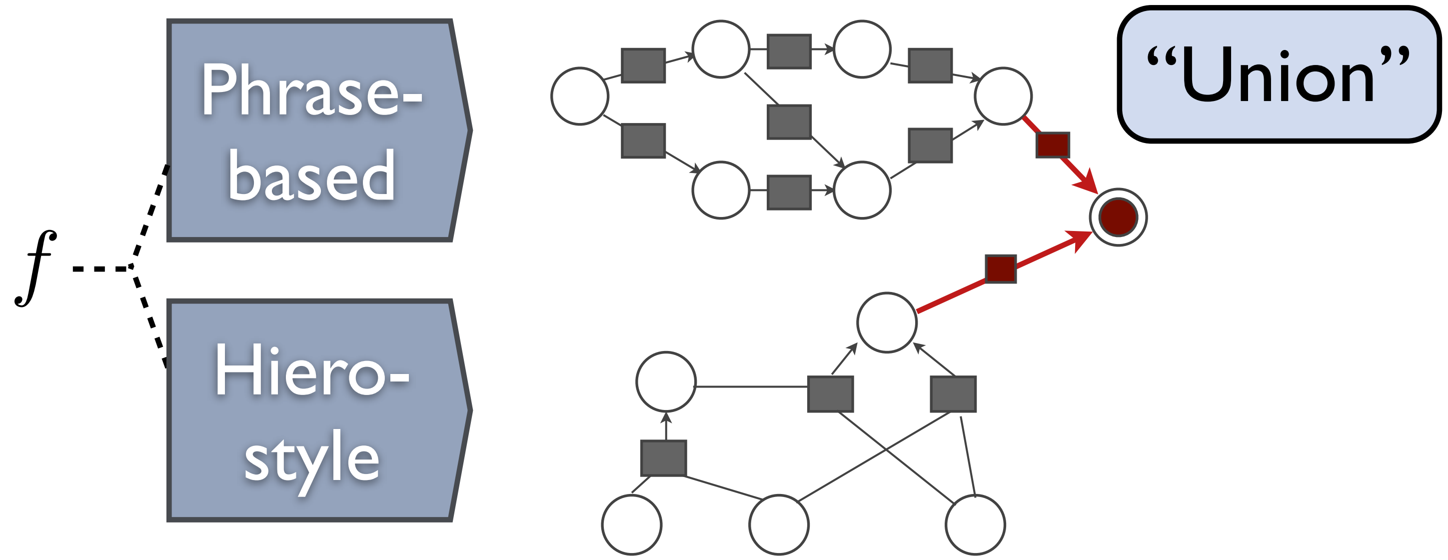
- Do we need model choice features?
- Should n-gram statistics be model specific?

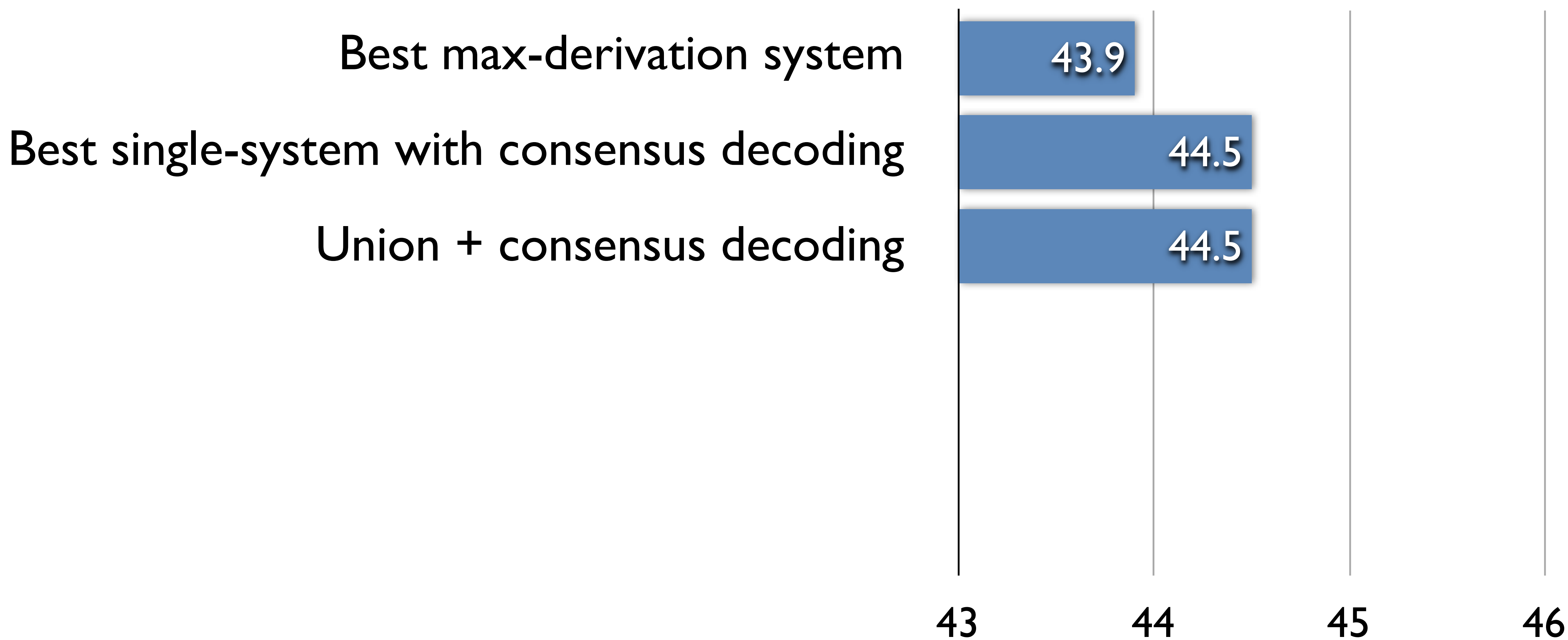




Sources of Improvement: Arabic-to-English

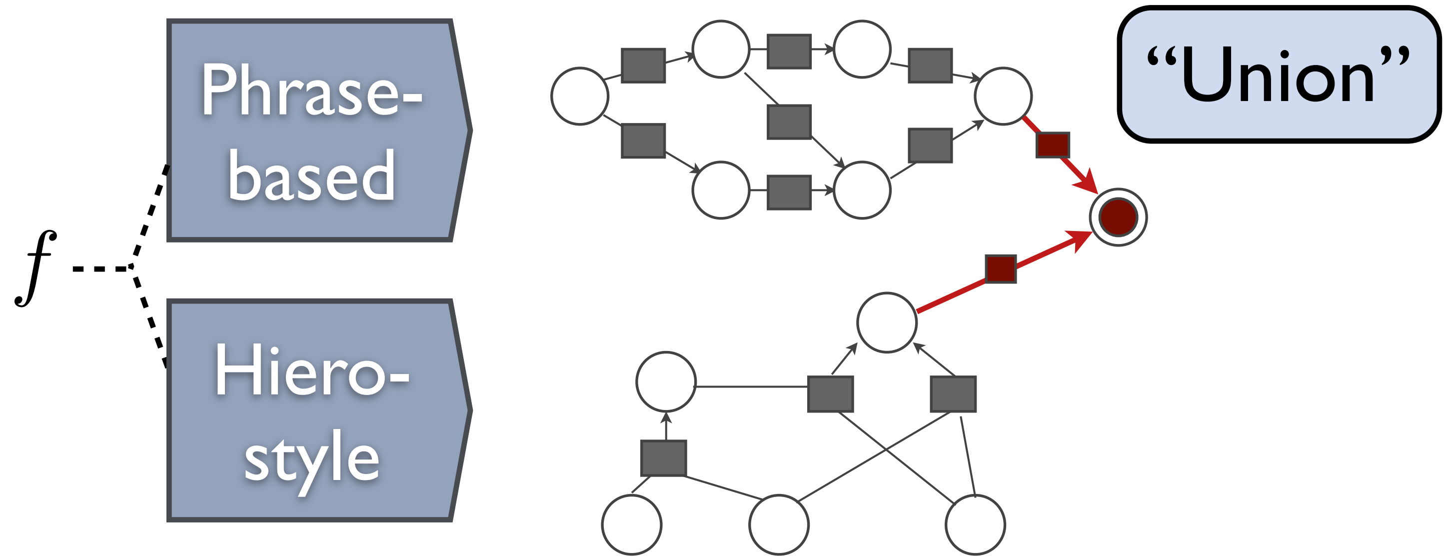
- Do we need model choice features?
- Should n-gram statistics be model specific?

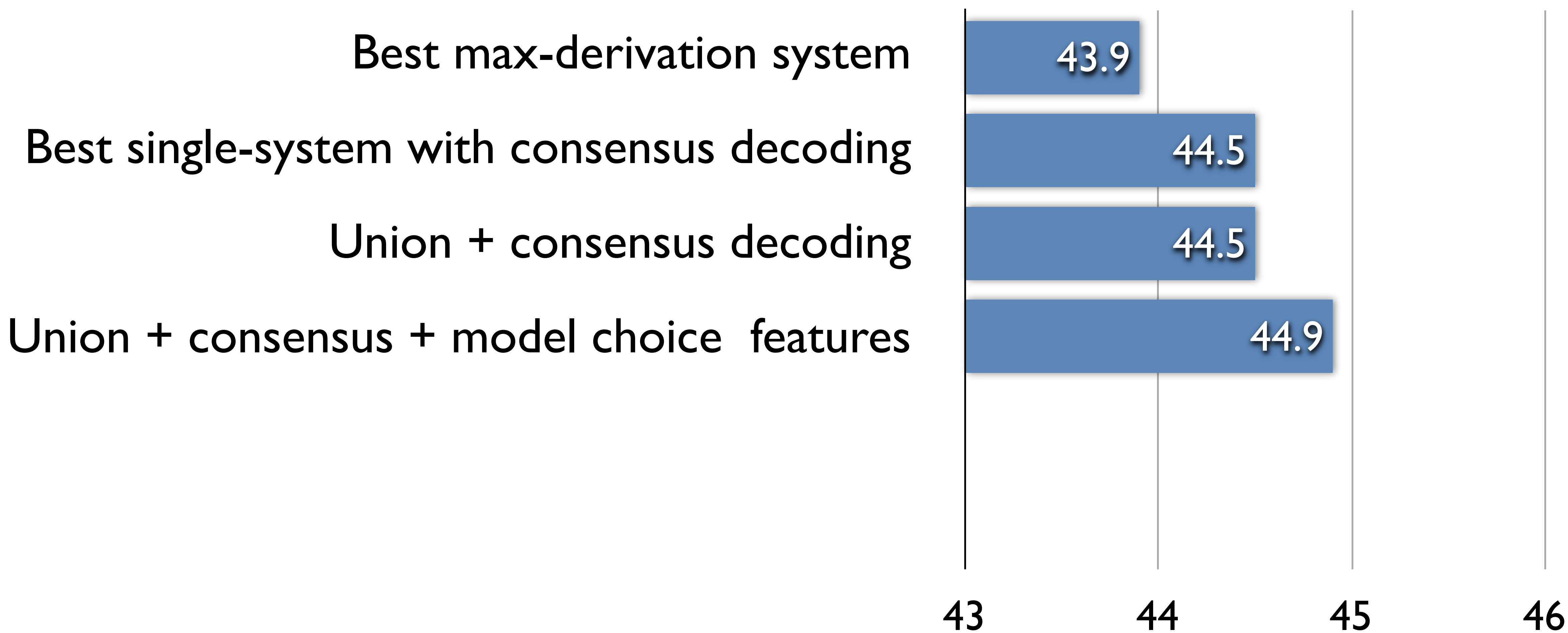




Sources of Improvement: Arabic-to-English

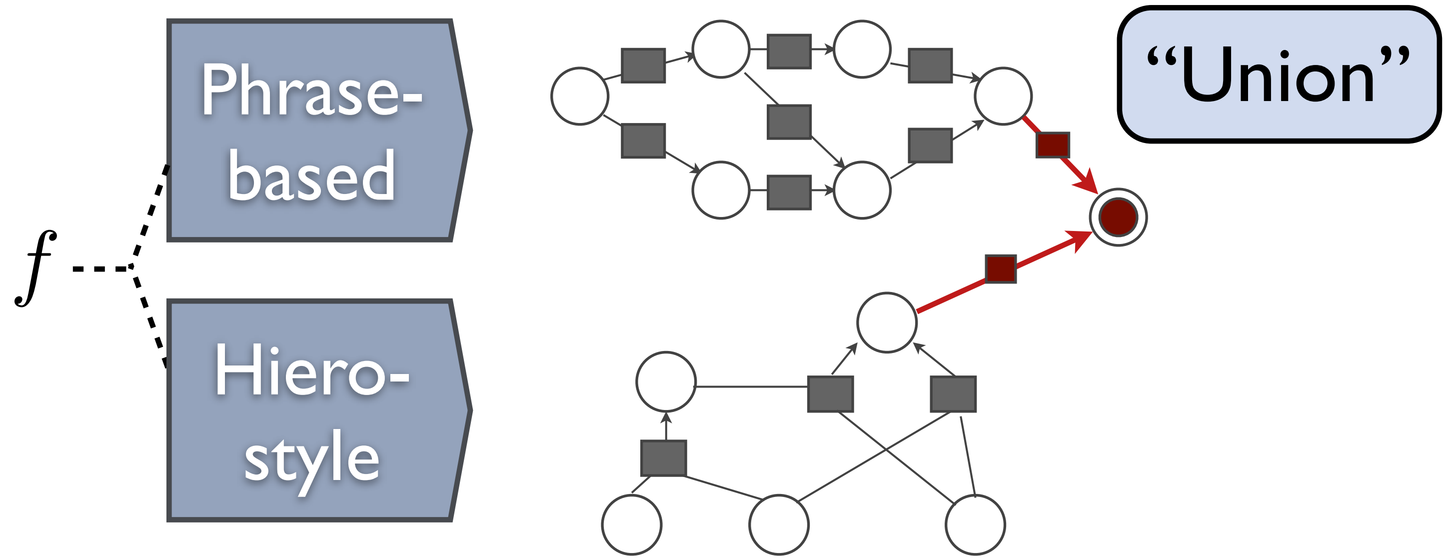
- Do we need model choice features?
- Should n-gram statistics be model specific?

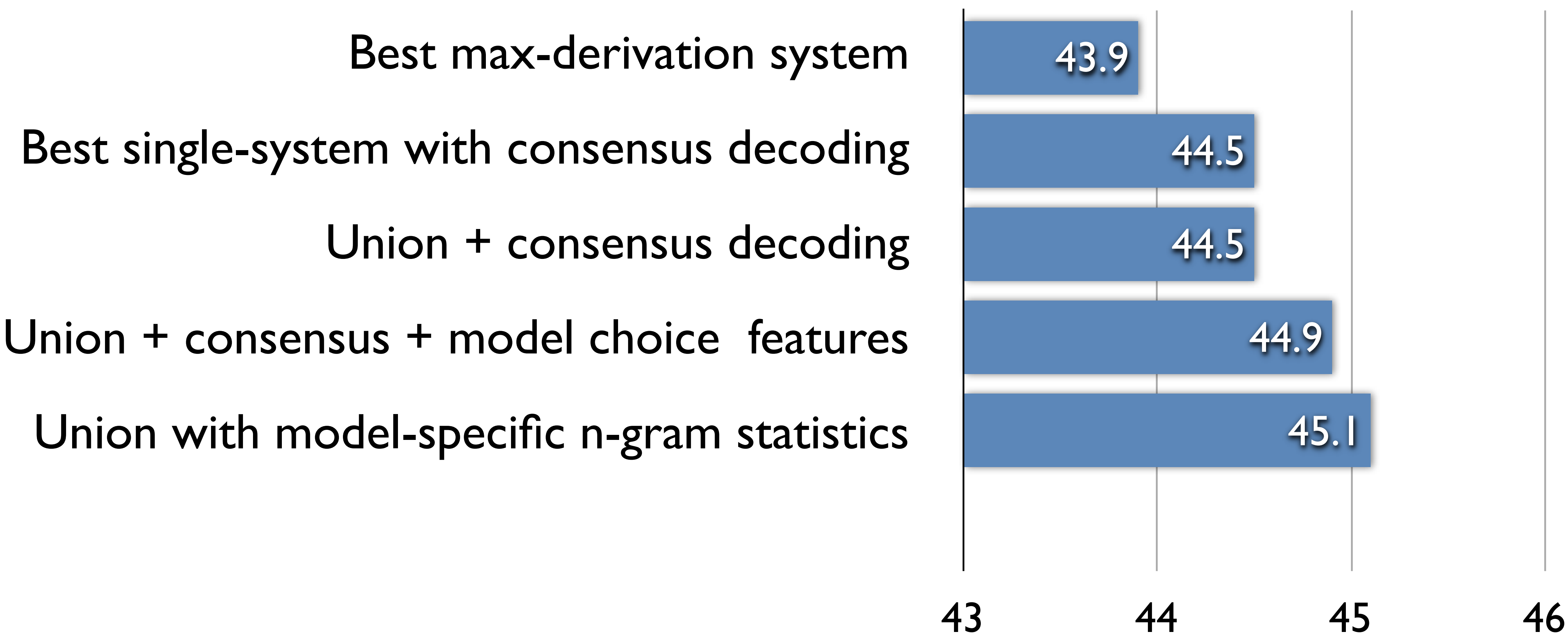




Sources of Improvement: Arabic-to-English

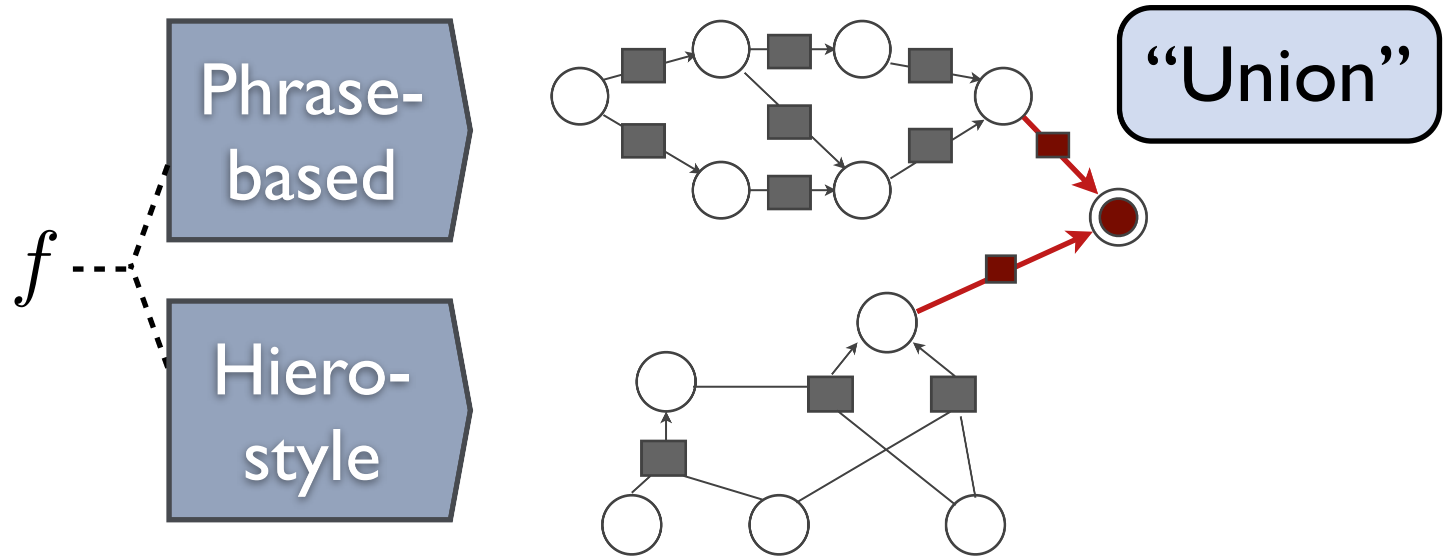
- Do we need model choice features?
- Should n-gram statistics be model specific?

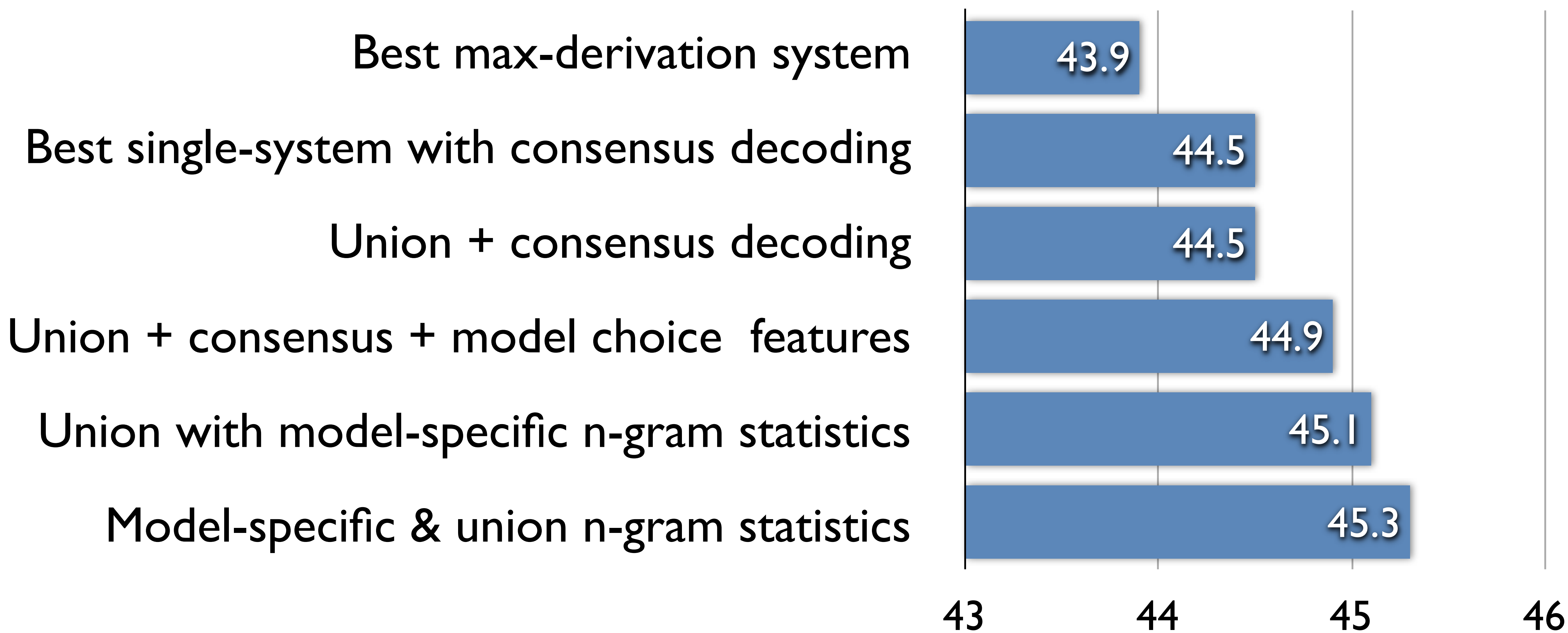




Sources of Improvement: Arabic-to-English

- Do we need model choice features?
- Should n-gram statistics be model specific?





Outline

Consensus decoding review

Our model combination technique

Comparison to system combination

Outline

Consensus decoding review

Our model combination technique

Comparison to system combination

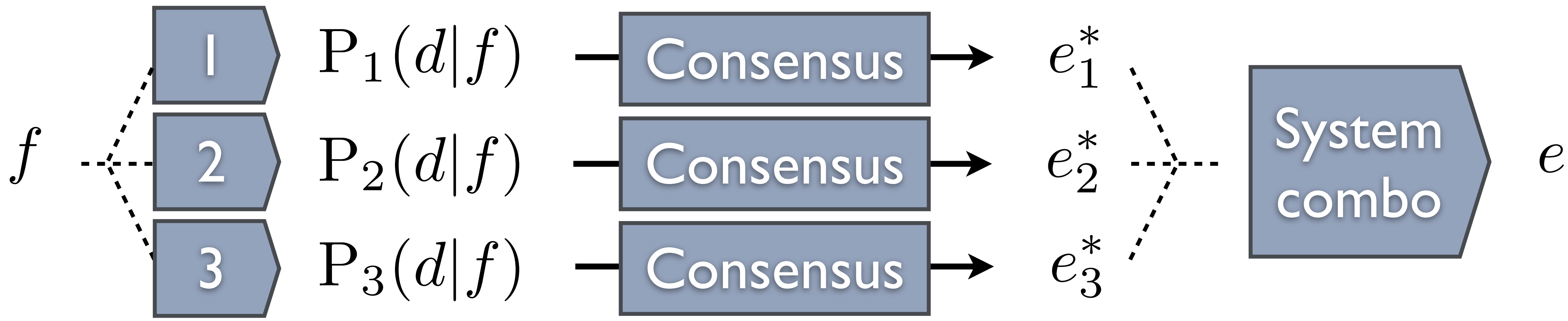
The Final Showdown

System Combination Baselines

Two established system combination methods [Macherey & Och, ’07]

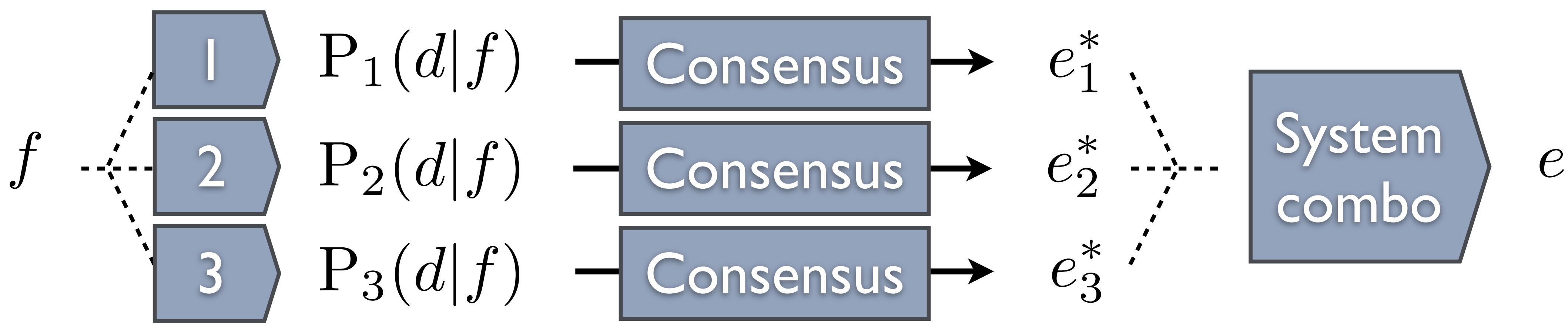
System Combination Baselines

Two established system combination methods [Macherey & Och, '07]



System Combination Baselines

Two established system combination methods [Macherey & Och, '07]

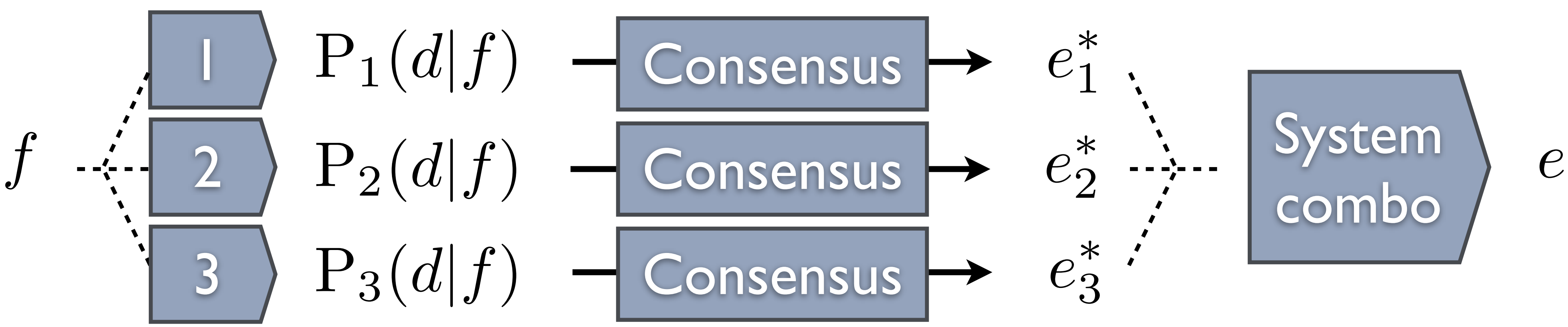


Sentence-level Choose among system outputs using an MBR objective

$$e = \arg\max_{e \in \{e_1^*, \ldots, e_k^*\}} \mathbb{E}\left[\mathrm{BLEU}(e)\right]$$

System Combination Baselines

Two established system combination methods [Macherey & Och, '07]



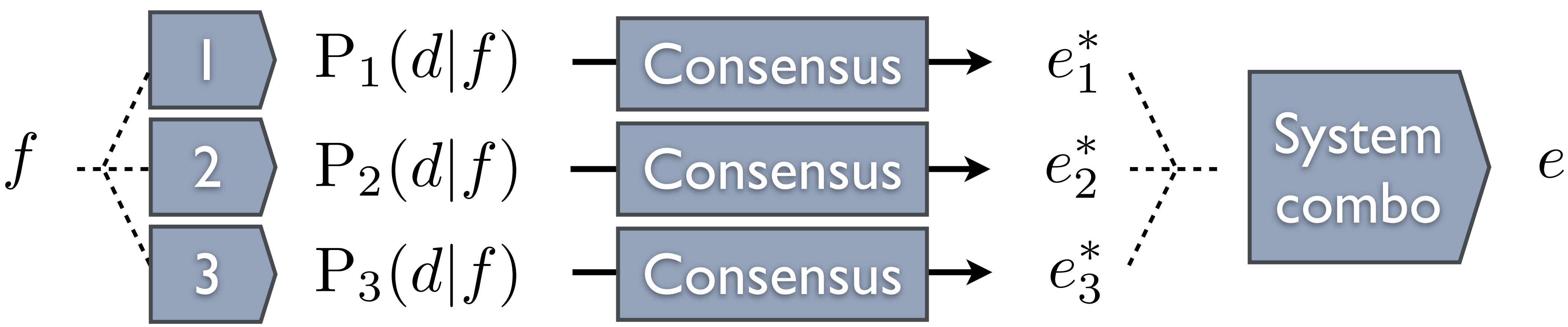
Sentence-level Choose among system outputs using an MBR objective

$$e = \underset{e \in \{e_1^*, \ldots, e_k^*\}}{\arg\max} \; \mathbb{E}\left[\mathrm{BLEU}(e)\right]$$

Word-level Confusion network approach

System Combination Baselines

Two established system combination methods [Macherey & Och, '07]



Sentence-level Choose among system outputs using an MBR objective

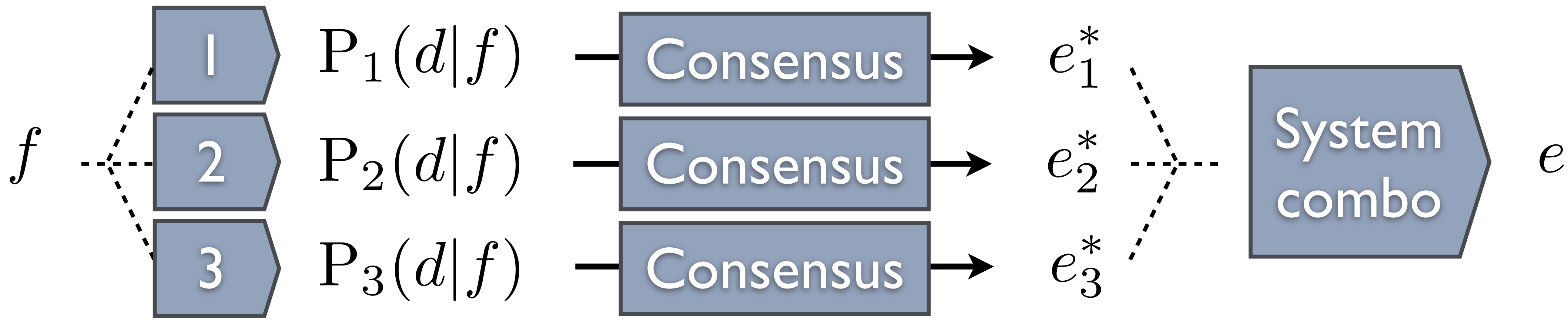
$$e = \arg\max_{e \in \{e_1^*, \ldots, e_k^*\}} \mathbb{E}\left[\mathrm{BLEU}(e)\right]$$

Word-level Confusion network approach

- All e_i^* are aligned to a backbone $e_b \in \{e_1^*, \ldots, e_k^*\}$

System Combination Baselines

Two established system combination methods [Macherey & Och, '07]



Sentence-level Choose among system outputs using an MBR objective

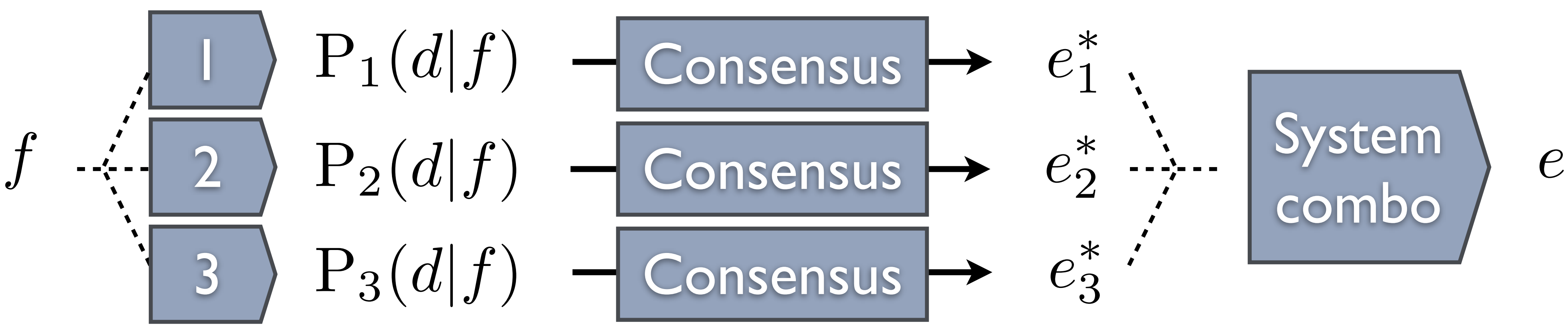
$$e = \arg\max_{e \in \{e_1^*, \ldots, e_k^*\}} \mathbb{E}\left[\text{BLEU}(e)\right]$$

Word-level Confusion network approach

- All e_i^* are aligned to a backbone $e_b \in \{e_1^*, \ldots, e_k^*\}$
- Sentences + alignments form a confusion network

System Combination Baselines

Two established system combination methods [Macherey & Och, '07]



Sentence-level Choose among system outputs using an MBR objective

$$e = \arg\max_{e \in \{e_1^*, \ldots, e_k^*\}} \mathbb{E}\left[\mathrm{BLEU}(e)\right]$$

Word-level Confusion network approach

- All e_i^* are aligned to a backbone $e_b \in \{e_1^*, \ldots, e_k^*\}$
- Sentences + alignments form a confusion network
- Output maximizes a consensus objective

Model versus System Combination

Gooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooogle

Qualitative

Quantitative

Model versus System Combination

Qualitative

- Single-system n-gram statistics are required in both methods

Quantitative

Model versus System Combination

Goooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooooogle

Qualitative

- Single-system n-gram statistics are required in both methods
- Model combination searches only once under a consensus objective

Quantitative

Model versus System Combination

Qualitative

- Single-system n-gram statistics are required in both methods
- Model combination searches only once under a consensus objective
- Inter-hypothesis alignment problem exists only in system combination

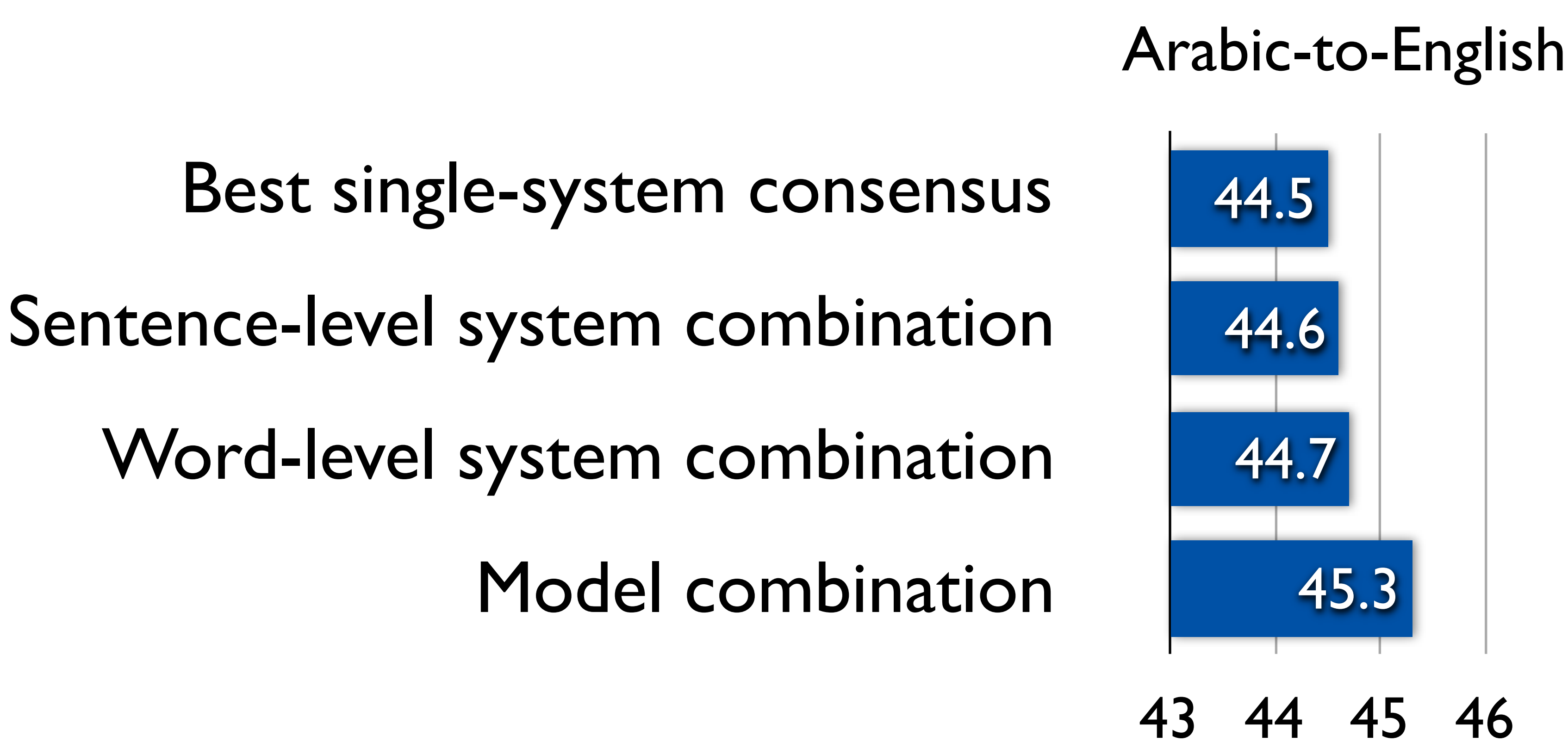
Quantitative

Model versus System Combination

Qualitative

- Single-system n-gram statistics are required in both methods
- Model combination searches only once under a consensus objective
- Inter-hypothesis alignment problem exists only in system combination

Quantitative

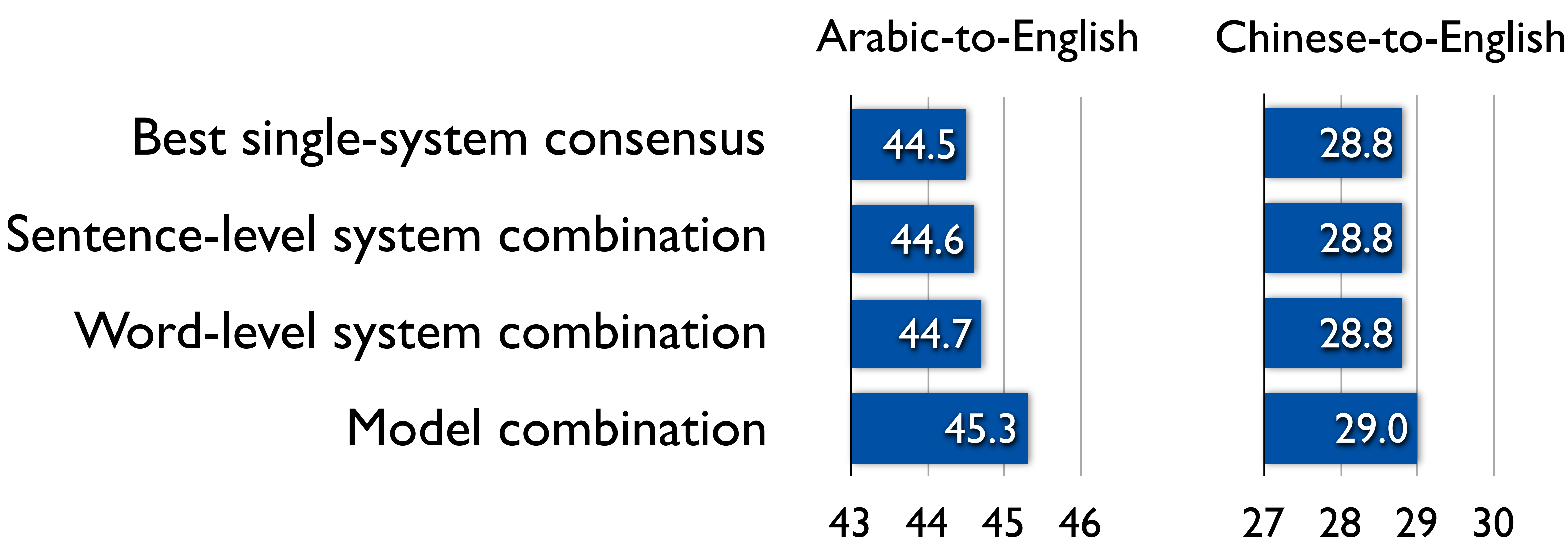


Model versus System Combination

Qualitative

- Single-system n-gram statistics are required in both methods
- Model combination searches only once under a consensus objective
- Inter-hypothesis alignment problem exists only in system combination

Quantitative



Conclusion

Conclusion

- Consensus decoding with a learned objective extends naturally to multiple models

Conclusion

- Consensus decoding with a learned objective extends naturally to multiple models

- Model combination provides consensus *and* combination effects in a unified, distribution-driven objective

Conclusion

- Consensus decoding with a learned objective extends naturally to multiple models

- Model combination provides consensus *and* combination effects in a unified, distribution-driven objective

- It outperforms a pipeline of consensus decoding followed by system combination, using less total computation

Conclusion

- Consensus decoding with a learned objective extends naturally to multiple models
- Model combination provides consensus *and* combination effects in a unified, distribution-driven objective
- It outperforms a pipeline of consensus decoding followed by system combination, using less total computation

It's easy, it's clean, and it works

Conclusion

- Consensus decoding with a learned objective extends naturally to multiple models

- Model combination provides consensus *and* combination effects in a unified, distribution-driven objective

- It outperforms a pipeline of consensus decoding followed by system combination, using less total computation

It's easy, it's clean, and it works

Thanks!