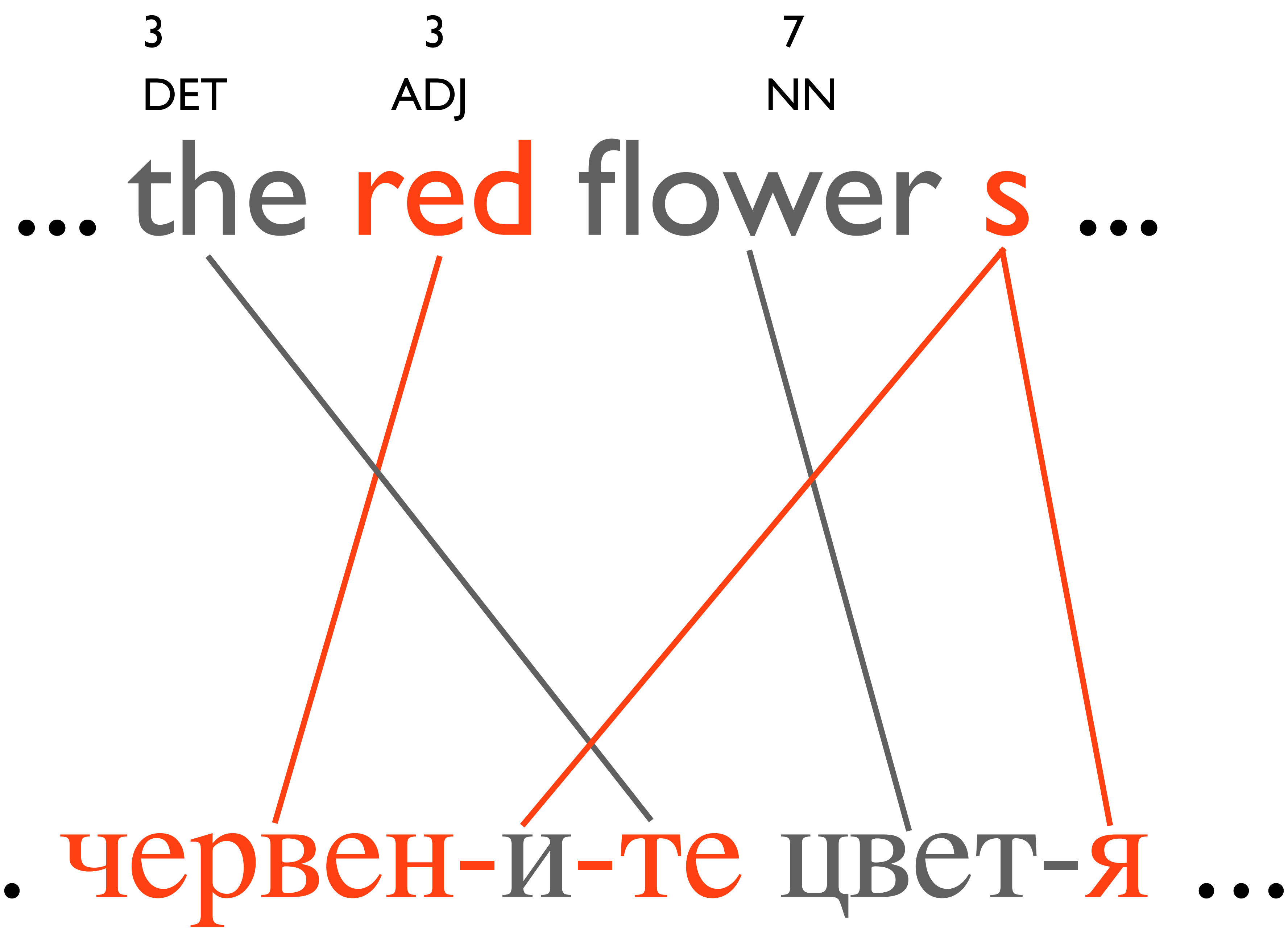
# Unsupervised Bilingual Morpheme Segmentation and Alignment

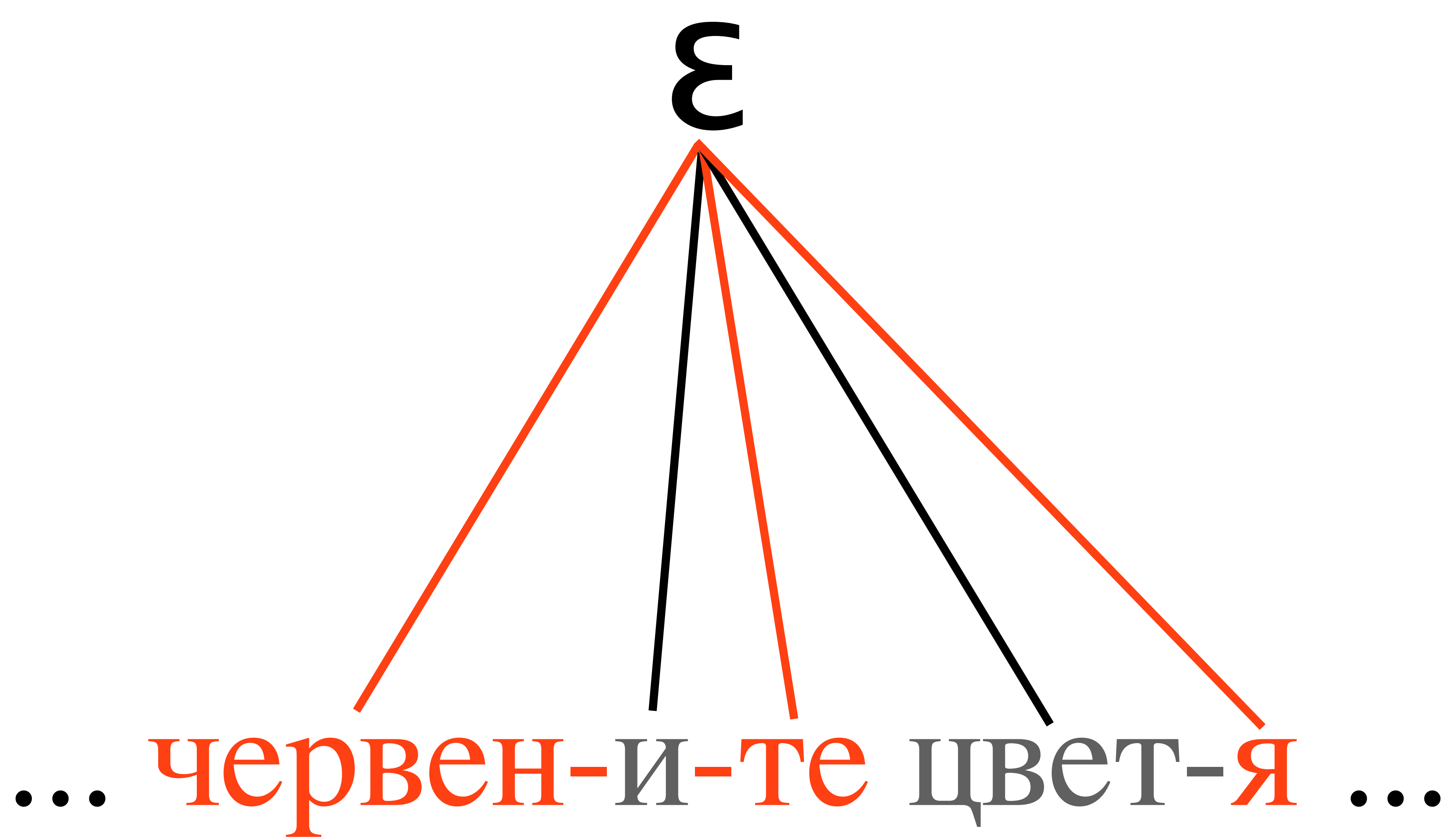
## ...with Context-rich Hidden Semi-Markov Models

Jason Naradowsky, UMass Amherst
Kristina Toutanova, Microsoft Research

# Context 1: Machine Translation



# Context 2: Segmentation



# Overview

- Motivation: why morphemes?
- Our Model
  - Preprocessing for Alignment
  - Model Components
  - Learning & Inference
- Experiments
  - Alignment
  - Segmentation

# Word Alignment

... the red flowers ...

... червените цветя ...

# Word Alignment

... the red flowers ...

... червените цветя ...

# Word Alignment

... the red flowers ...

... червените цветя ...

# Word Alignment

... the red flowers ...

... червените цветя ...

# Word Alignment

... (the red) flowers ...

... червените цветя ...

# Word Alignment

... (the red) flower-s ...

... червен-и-те цвет-я ...

# Morphological Productivity

the red flowers: червен-и-те
(plural)

# Morphological Productivity

the red flower: червен-о-то

(neuter)

# Morphological Productivity

the red book: червен-а-та

(feminine)

# Morphological Productivity

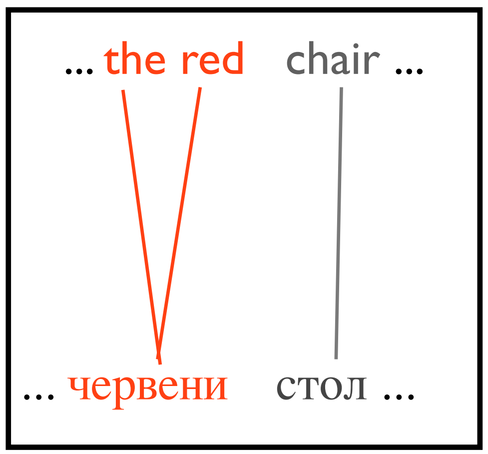
the red chair: червеният

(masculine)

# Morphological Productivity

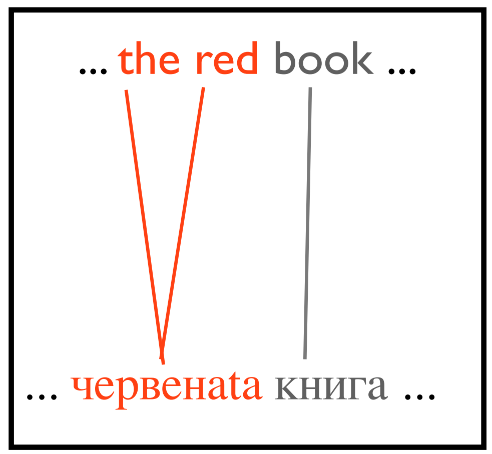
**masc.**

червеният | red



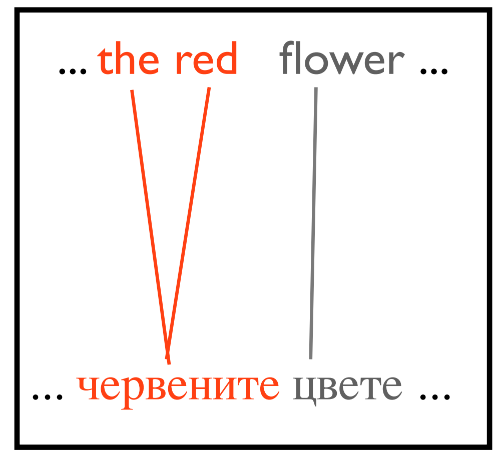

**fem.**

червената | red



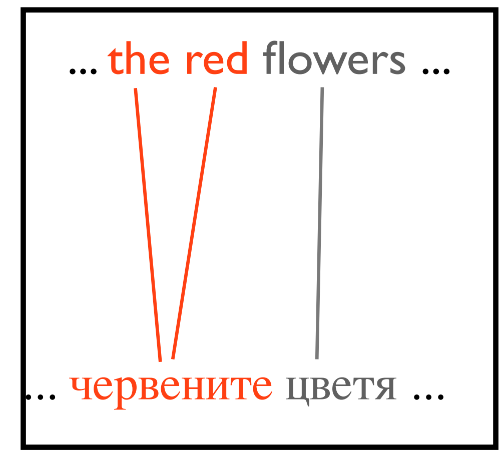

**neuter**

червеното | red




**plural**

червените | red



# Morphological Productivity

| | masc. | fem. | neuter | plural |
|---|---|---|---|---|
| | червеният \| red | червената \| red | червеното \| red | червените \| red |
| def. | ... the red chair ...<br>... червени стол ... | ... the red book ...<br>... червената книга ... | ... the red flower ...<br>... червениото цвете ... | ... the red flowers ...<br>... червените цветя ... |
| indef. | ... red chair ...<br>... червен стол ... | ... red book ...<br>... червена книга ... | ... red tree ...<br>... червено цвете ... | ... red flowers ...<br>... червени цвете ... |
| | червен \| red | червена \| red | червено \| red | червени \| red |

# Morphological Productivity Yields Sparsity!

| Counts | |
|---|---|
| червен | red | 1 |
| червената | red | 1 |
| червеното | red | 1 |
| червените | red | 1 |
| червеният | red | 1 |
| червена | red | 1 |
| червено | red | 1 |
| червени | red | 1 |

vs червен | red × 8

# Previous Work

- Snyder & Barzilay (2008)
    - A generative model for finding morphological paradigms across languages.
- Xu et al. (2008)
    - A Bayesian model for segmenting Chinese for use in MT.
- Chung & Gildea (2009)
    - Target tokenization and alignment for MT using IBM Model 1 assumptions.

Task: Resource-Rich → Resource-Poor Translation

Goal: Find best target segmentation and alignment to source morphemes.

- Our Contributions
  - Hidden semi-markov model to find target morpheme segmentation
  - Leverage source-side information
  - Broad contextual dependencies with hierarchical smoothing
  - Latent morphological type induction

# Overview

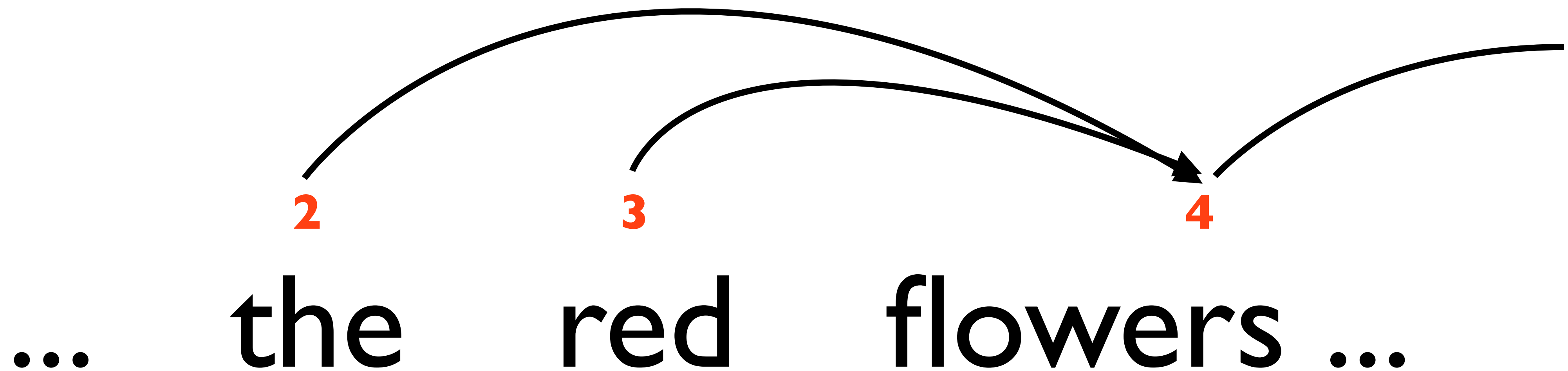
- Motivation: why morphemes?
- **Our Model**
  - Preprocessing for Alignment
  - Model Components
  - Learning & Inference
- Experiments
  - Alignment
  - Segmentation

# Marking up the source side:

... the red flowers ...

# Marking up the source side:

1. Tag
2. Parse




| | | | |
|---|---|---|---|
| **HEAD:** | 4 | 4 | 7 |
| **POS:** | DET | ADJ | NN |

# Marking up the source side:

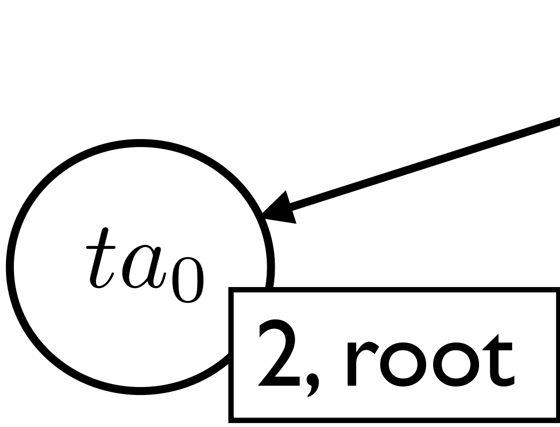
1. Tag
2. Parse
3. Segment

| | | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| | … | the | red | flower | s … |
| **HEAD:** | | 4 | 4 | 7 | 7 |
| **POS:** | | DET | ADJ | NN | NN-SUFFIX |

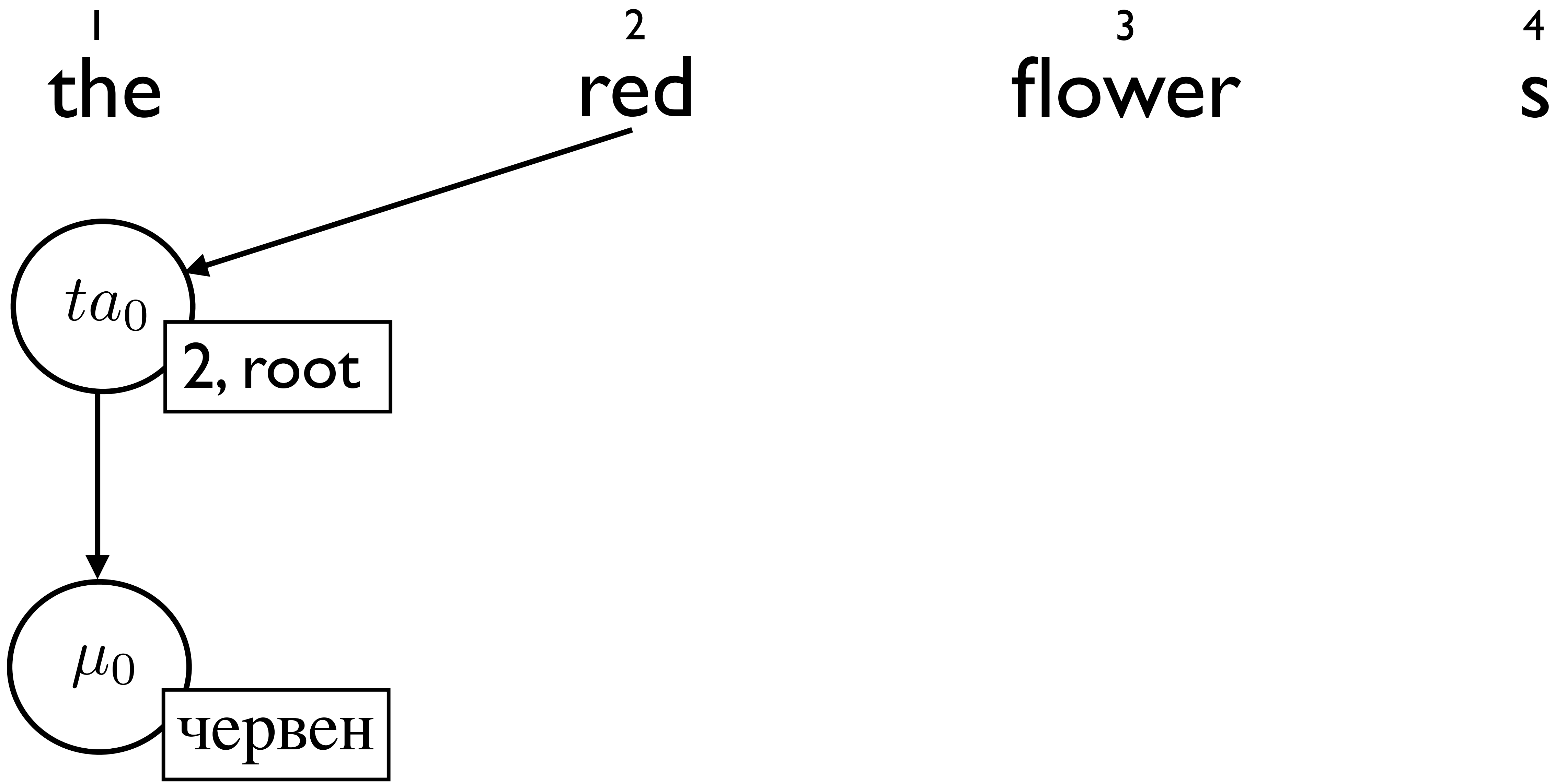
# Overview

- Motivation: why morphemes?
- **Our Model**
  - Preprocessing for Alignment
  - **Model Components**
  - Learning & Inference
- Experiments
  - Alignment
  - Segmentation

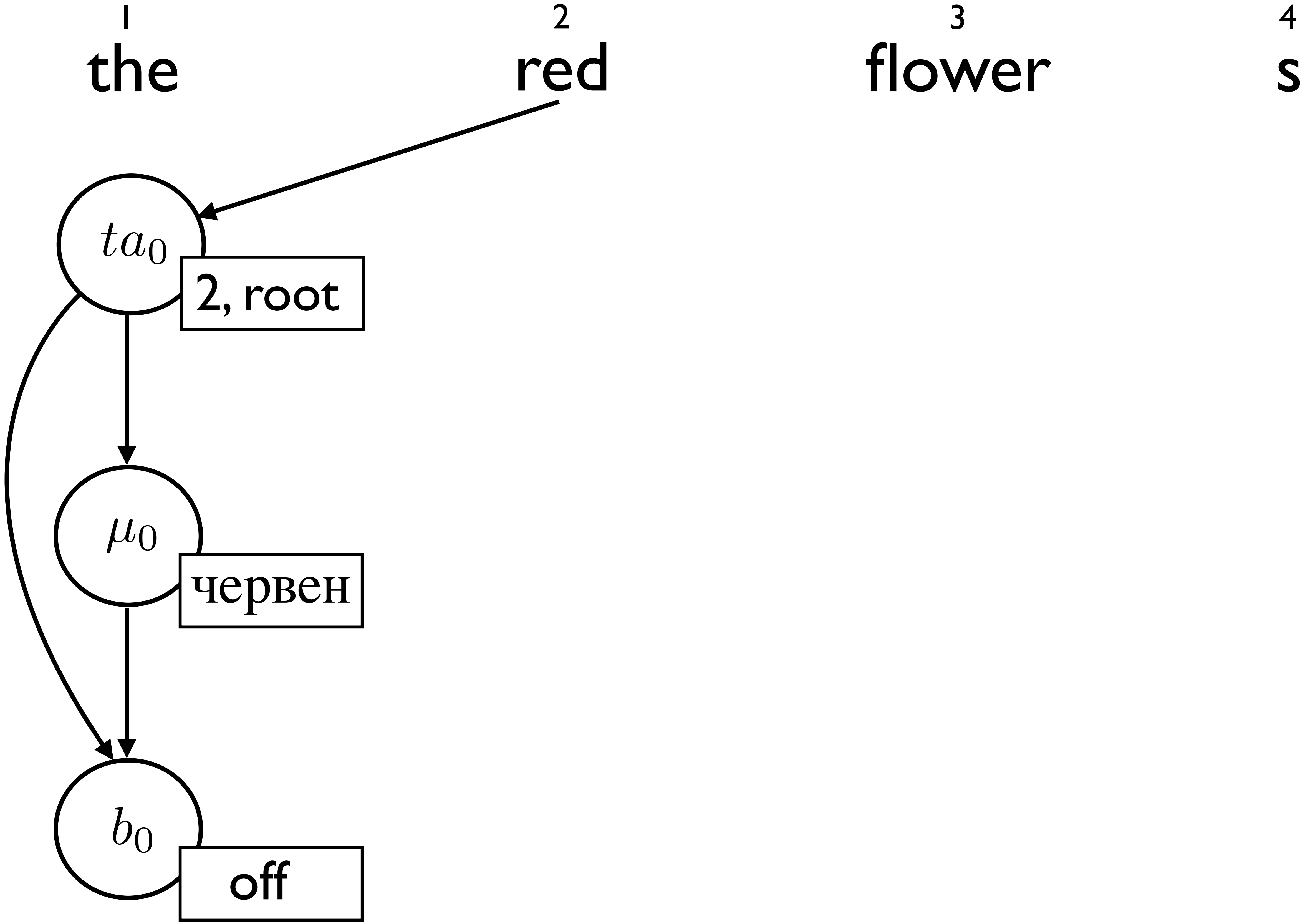
| 1 | 2 | 3 | 4 |
|---|---|---|---|
| the | red | flower | s |



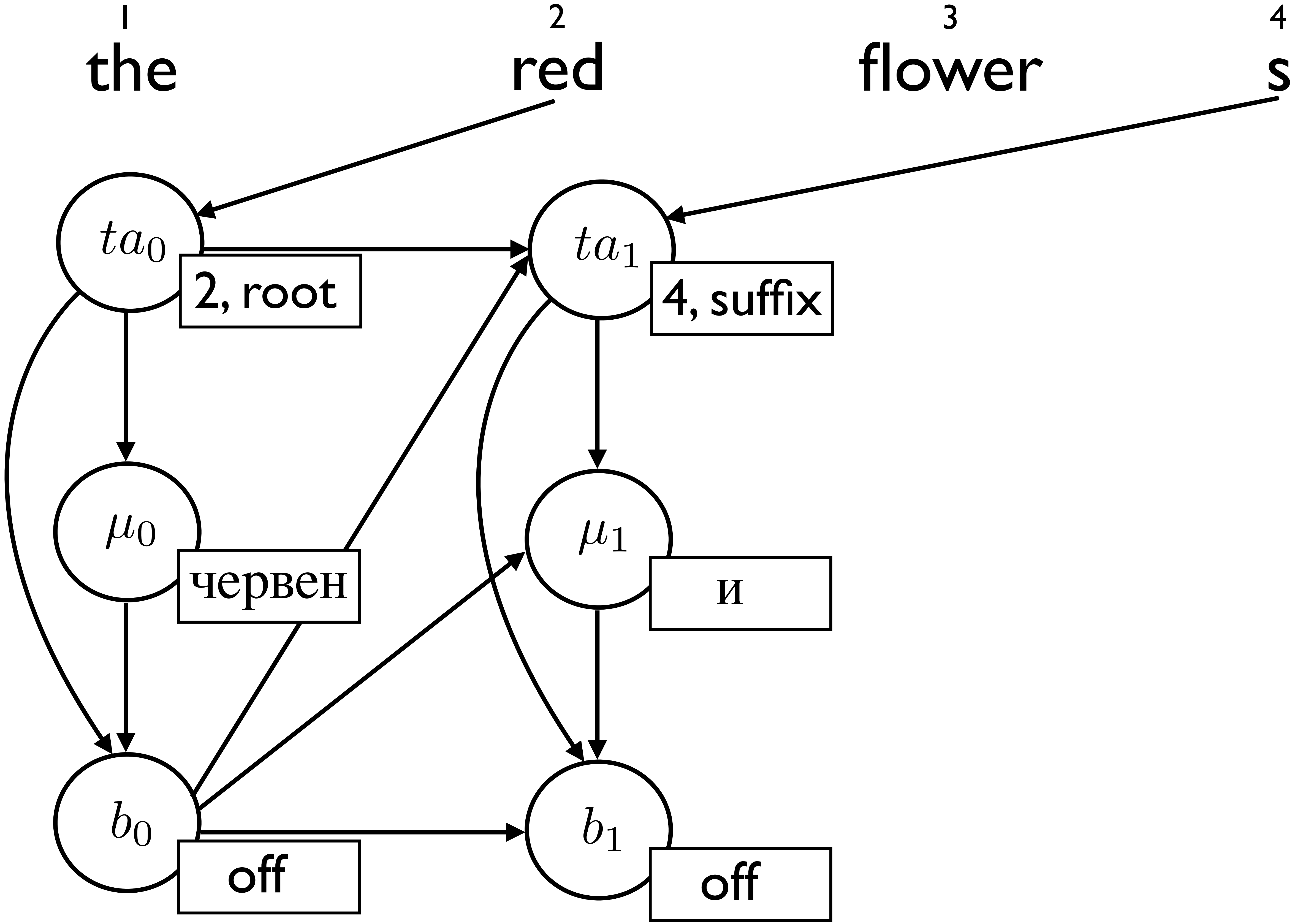

... червен-и-те ...



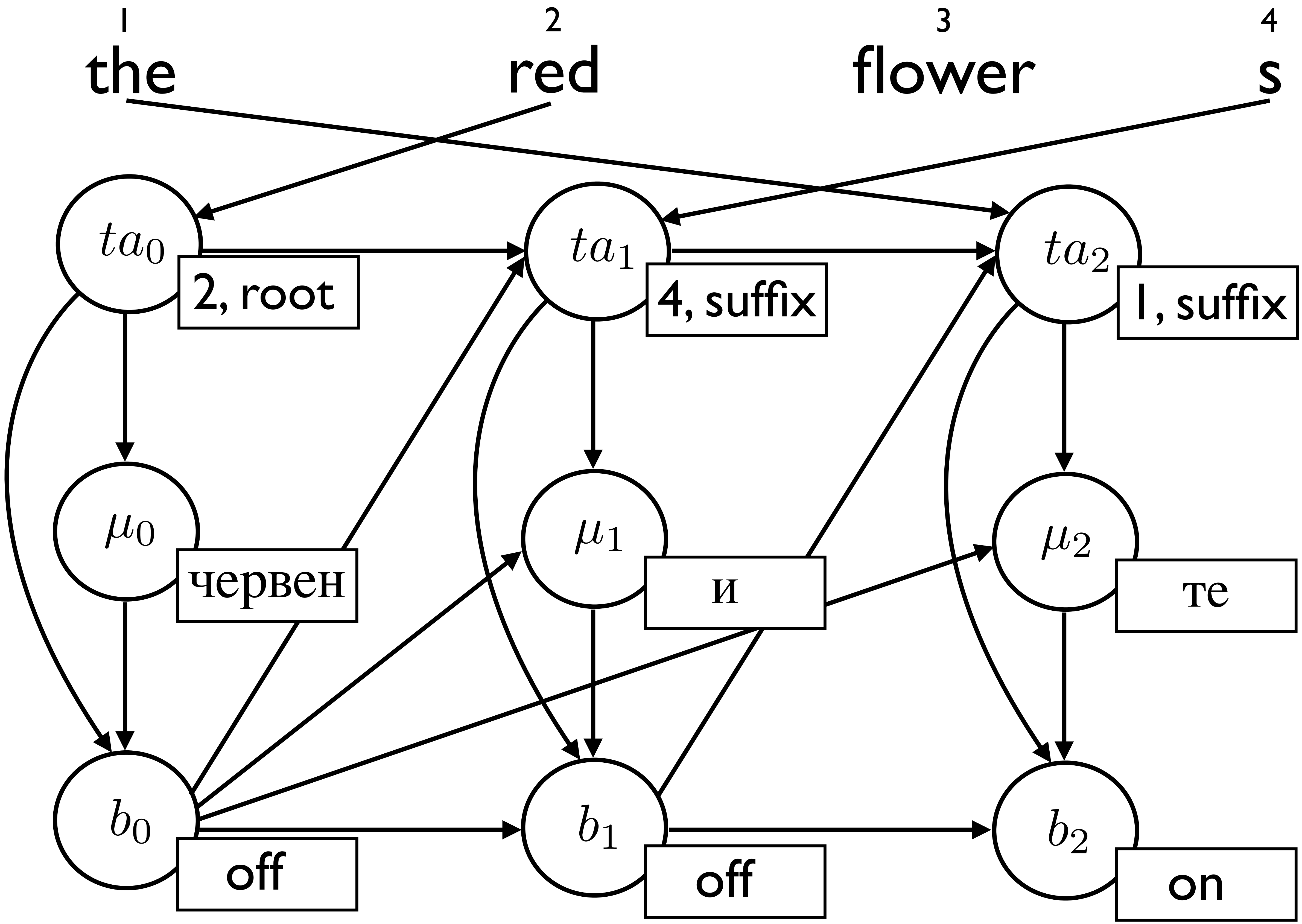

... червен-и-те ...




... червен-и-те ...


1
the
2
red
3
flower
4
s
$ta_0$
2, root
$ta_1$
4, suffix
$\mu_0$
червен
$\mu_1$
и
$b_0$
off
$b_1$
off

... червен-и-те ...


1
2
3
4
the
red
flower
s
$ta_0$
2, root
$ta_1$
4, suffix
$ta_2$
1, suffix
$\mu_0$
червен
$\mu_1$
и
$\mu_2$
те
$b_0$
off
$b_1$
off
$b_2$
on

... червен-и-те ...

# Model Components

- Distortion Model
- Morpheme Translation Model
- Word Boundary Generation Model
- Length Penalty

# Model Decomposition:

distortion model

$$P(\mu, \mathbf{ta}, \mathbf{b}|\mathbf{e}) = \prod_{i=1}^{I} P_D(ta_i|ta_{i-1}, b_{i-1}, \mathbf{e})$$

morpheme translation model

$$\cdot P_T(\mu_i|ta_i, b_{i-1}, b_{i-2}, \mu_{i-1}, \mathbf{e})$$

word boundary model

$$\cdot P_B(b_i|\mu_i, \mu_{i-1}, ta_i, b_{i-1}, b_{i-2}, \mathbf{e})$$

length penalty

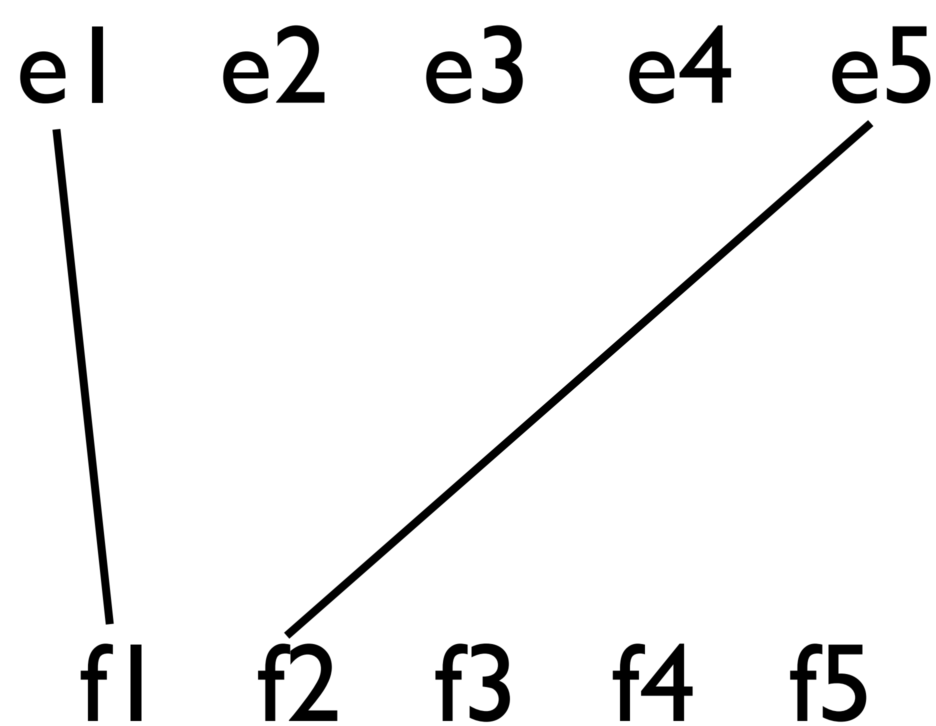
$$\cdot LP(|\mu_i|)$$

# Distortion Model

## Traditional Form:

$$P_D = P(a_i | a_{i-1}, \mathbf{e})$$

## Motivation:

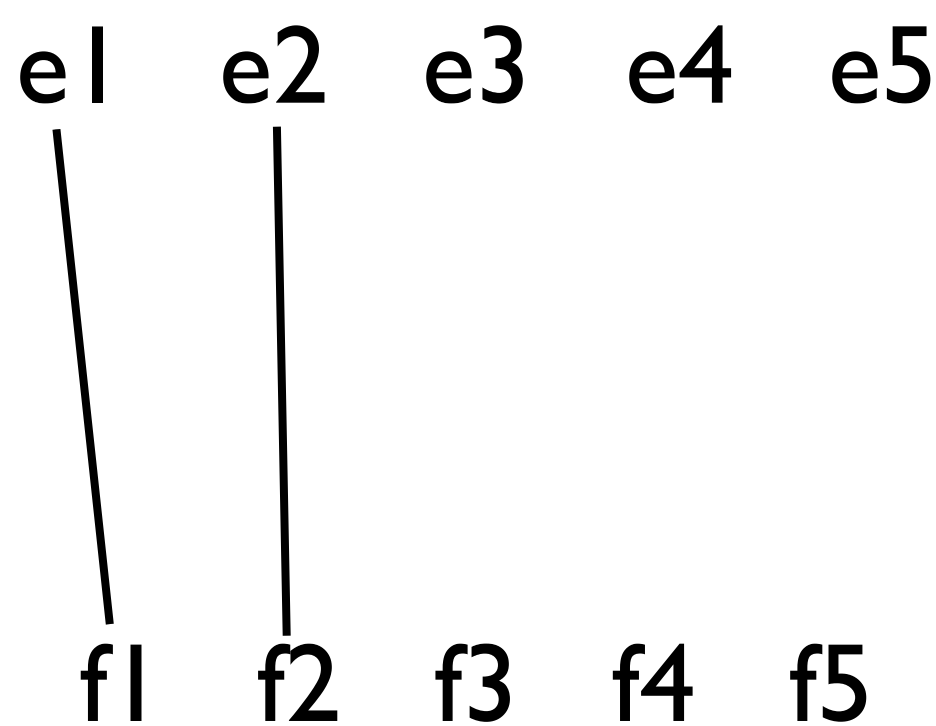


# Distortion Model

## Traditional Form:

$$P_D = P(a_i | a_{i-1}, \mathbf{e})$$

## Motivation:



# Distortion Model

Traditional form:

$$P_D = P(a_i | a_{i-1}, \mathbf{e})$$

Replaced with log-linear model:

$$P_D = \frac{e^{\phi(a_i, a_{i-1}, \mathbf{e})}}{\sum_i e^{\phi(a_i, a_{i-1}, \mathbf{e})}}$$

(Berg-Kirkpatrick 2010)

# Distortion Model

Actual form:

$$P_D(ta_i | ta_{i-1}, b_{i-1}, \mathbf{e})$$

- Richer context
- $t$ variables capture morphological type, {prefix, root, suffix}

# Distortion Features

| Feature | Value |
|---|---|
| Morph Distance | 1 |
| Word Distance | 1 |
| Binned Morph Distance | fore1 |
| Binned Word Distance | fore1 |
| Morph State Transition | suffix-root |
| Same Target Word | FALSE |
| POS Tag Transition | DET-NN |
| Dep Relation | DET←NN |
| Null Alignment | FALSE |

... and conjunctions

- Morpheme Translation Model

$$P_T(\mu_i | ta_i, b_{i-1}, b_{i-2}, \mu_{i-1}, \mathbf{e})$$

  - Also depend on aligned source morpheme and POS

- Hierarchical Back-off:

$$P_T(\mu_i | e_{a_i}, t_i) = \frac{c(\mu_i, e_{a_i}, t_i) + \alpha_2 P_2(\mu_i | t_i)}{c(e_{a_i}, t_i) + \alpha_2}$$

$$P_2(\mu_i | t_i) = \frac{c(\mu_i, t_i) + \alpha_i P_1(\mu_i)}{c(t_i) + \alpha_i}$$

$$P_1(\mu_i) = \frac{c(\mu_i) + \alpha_0 P_0(\mu_i)}{c(.) + \alpha_0}$$

# Word Boundary Generation

$$\cdot P_B(b_i|\mu_i, \mu_{i-1}, ta_i, b_{i-1}, b_{i-2}, \mathbf{e})$$

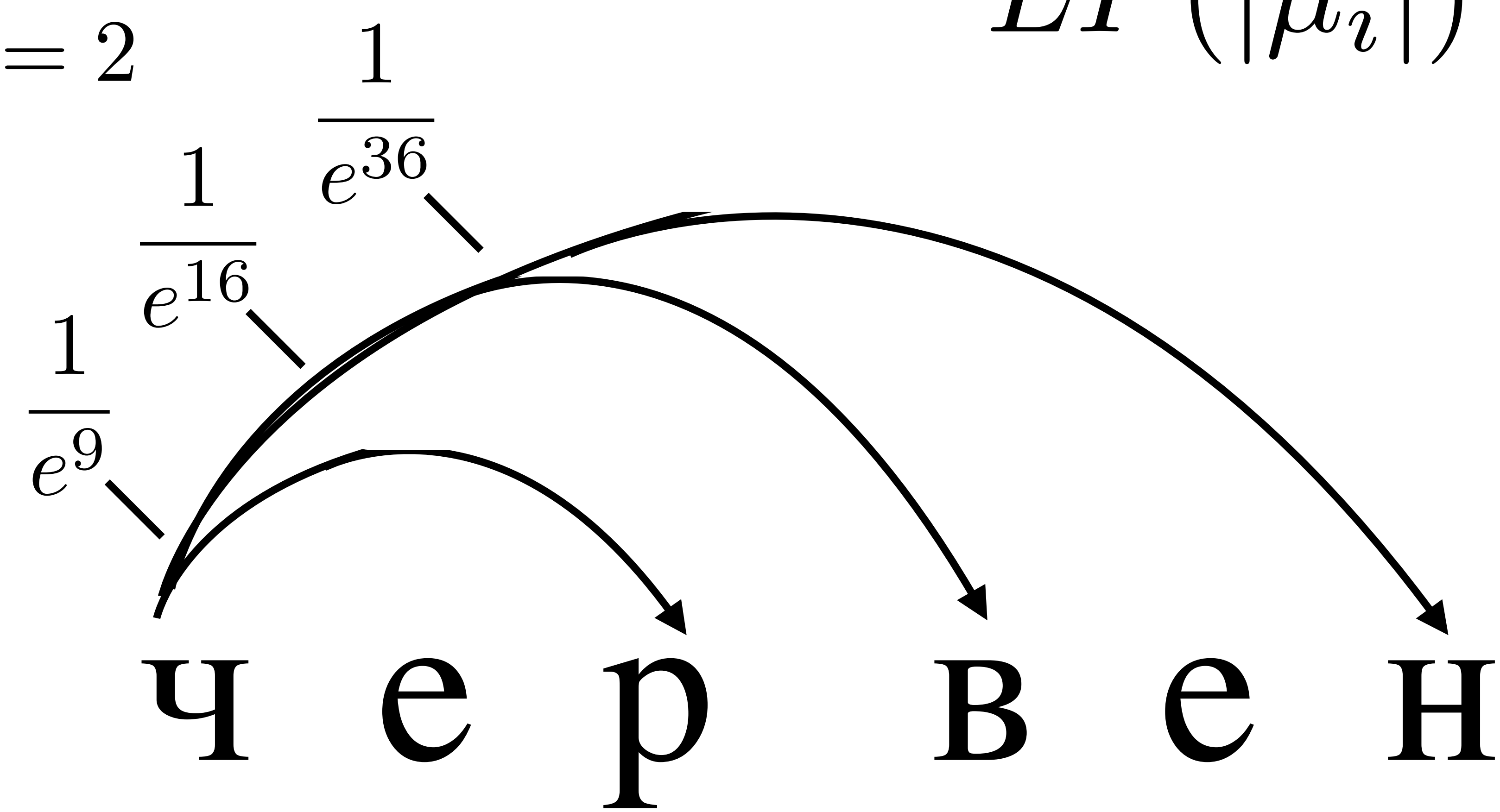
- Useful as contextual information
  - (Poon 09, Creutz & Lagus 07)
- Explicitly expressed in the model
  - Estimate what morphemes are likely to have which position in a word, number of morphemes in a word
- Observed on target side - no increase to inference complexity

# Length Penalty:

(Chung & Gildea 09),
(Liang & Klein 09)

$$LP(|\mu_i|) = \frac{1}{e^{|\mu_i|^{l_p}}}$$

$$ex: l_p = 2$$






Thursday, June 23, 2011

# Overview

- Motivation: why morphemes?
- **Our Model**
  - Preprocessing for Alignment
  - Model Components
  - **Learning & Inference**
- Experiments
  - Alignment
  - Segmentation

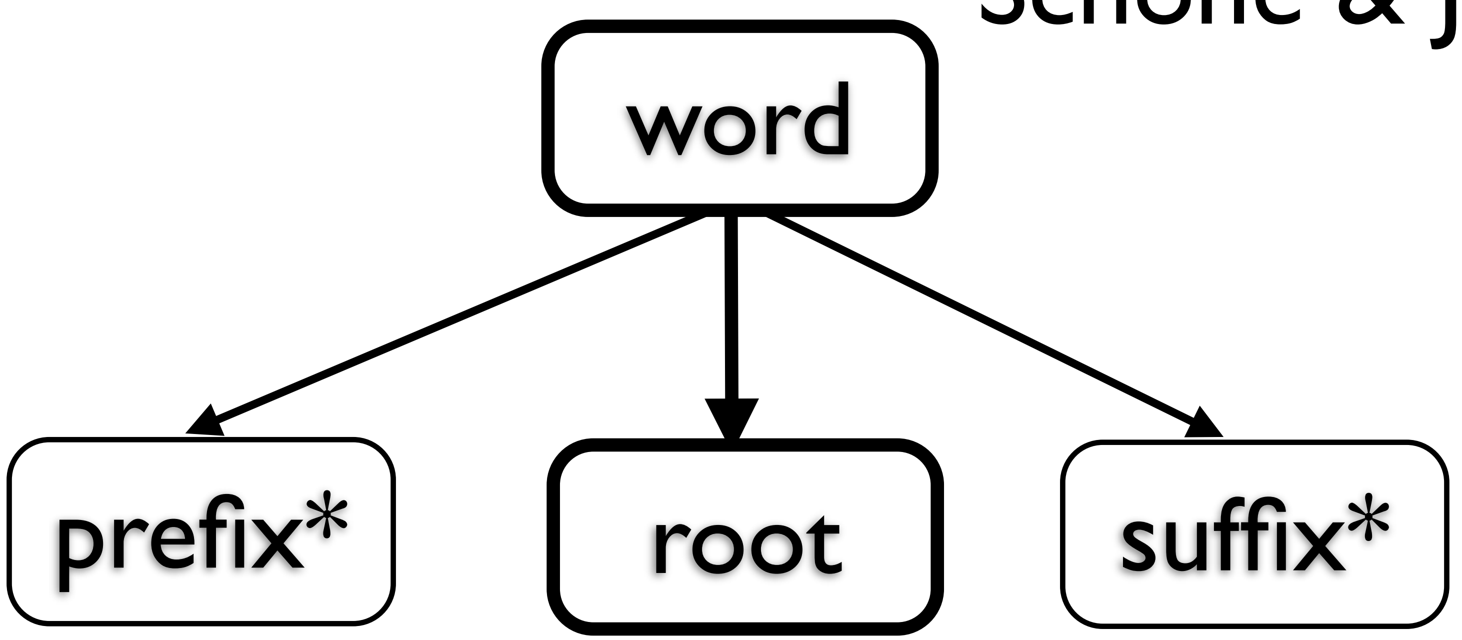
# Inference

- Exact, polynomial (standard semi-markov with richer state-space)
- EM E-Step:
  - Counts taken for valid hidden variable configurations
- EM M-Step
  - LBFGS for Distortion model
  - Interpolation counts for translation and word boundary model (a way of backing-off to less conditioning)

# Pruned Decoding

- Create a dictionary from the target corpus
  - Insert each word into a trie
- Derive list of top K most frequent affixes
- Restrict model to:

Schone & Jurafsky (2000)



# Overview

- Motivation: why morphemes?
- Our Model
    - Preprocessing for Alignment
    - Model Components
    - Learning & Inference
- **Experiments**
    - Alignment
    - Segmentation

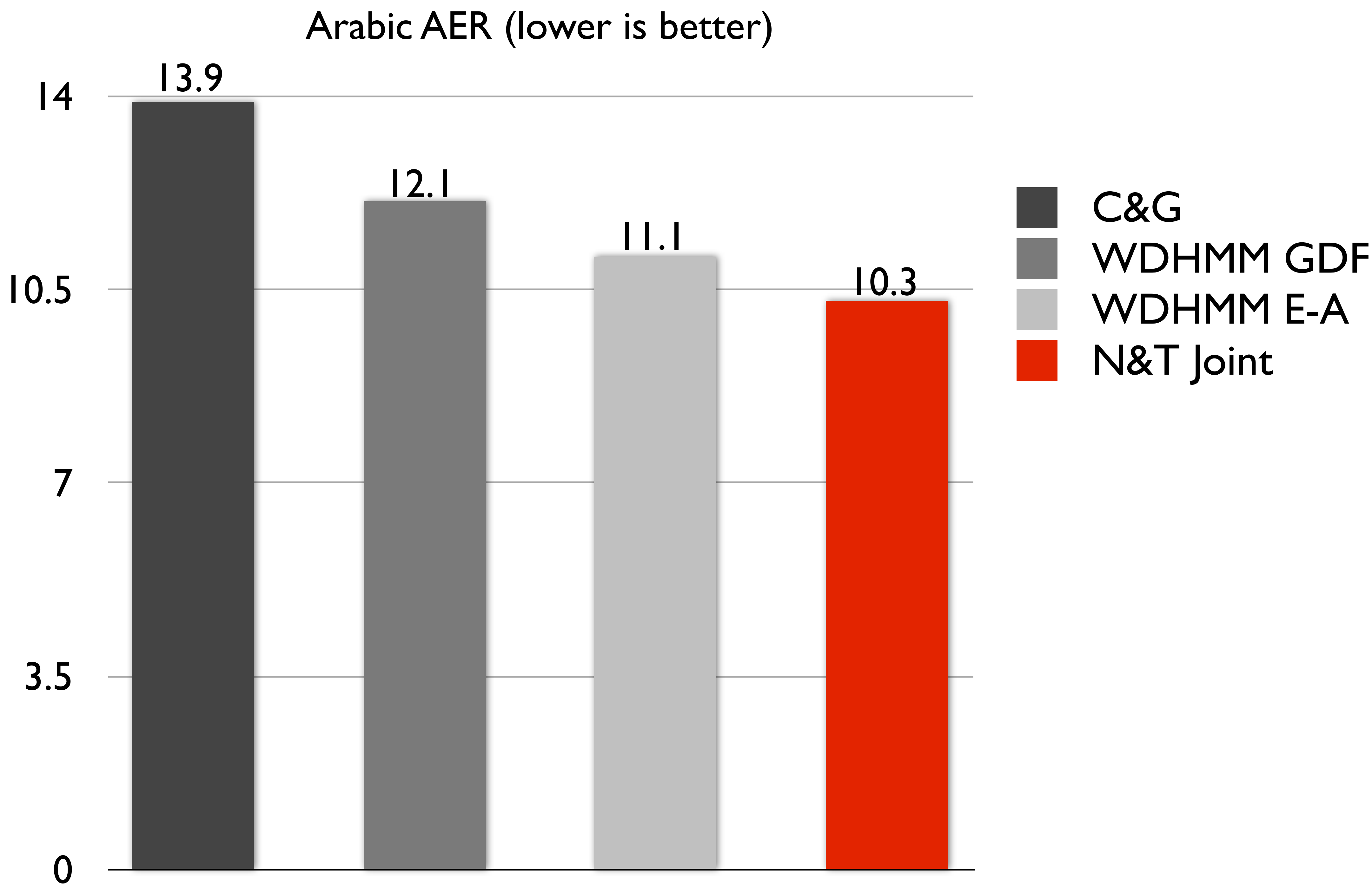
# Data

- Parallel phrases corpus (Snyder & Barzilay 2008)
  - 6,139 short phrases drawn from English, Hebrew, and Arabic Bible text.
  - Manually annotated Arabic with morpheme alignments
- Arabic Treebank
  - 140,265 words
- Both have gold morphological analyses
- Held-out data for smoothing, dictionary size, conditioning context, length penalties

# Alignment Experiments

- Procedure
  - Typical joint training
  - Project morpheme alignments to word alignment
- Evaluated on Alignment Error Rate (AER)
- Results are for Arabic short phrases only!
- Baseline 1: Chung & Gildea (2009)
- Baseline 2: WDHMM model of He (2007)

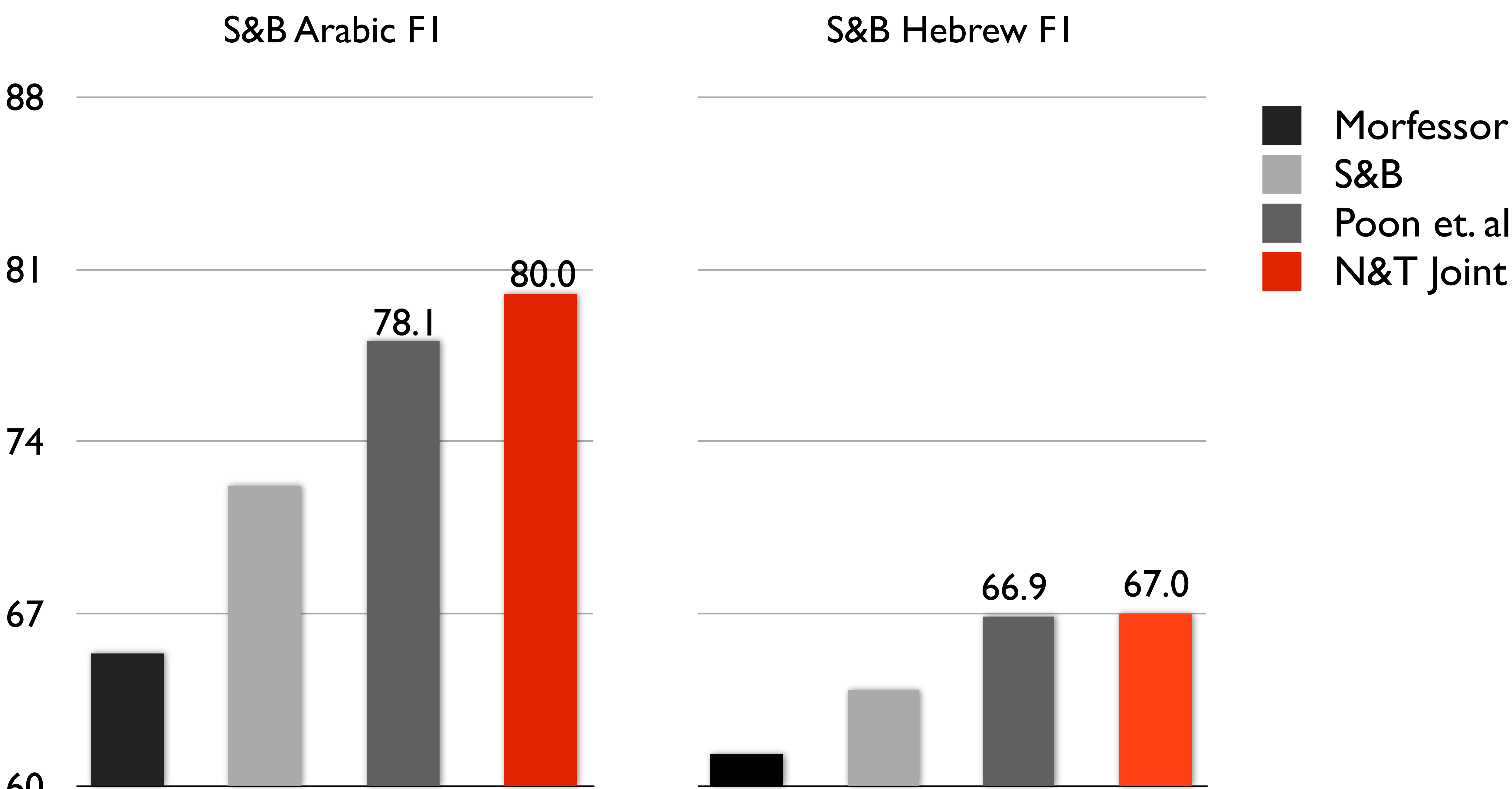
# Alignment Results



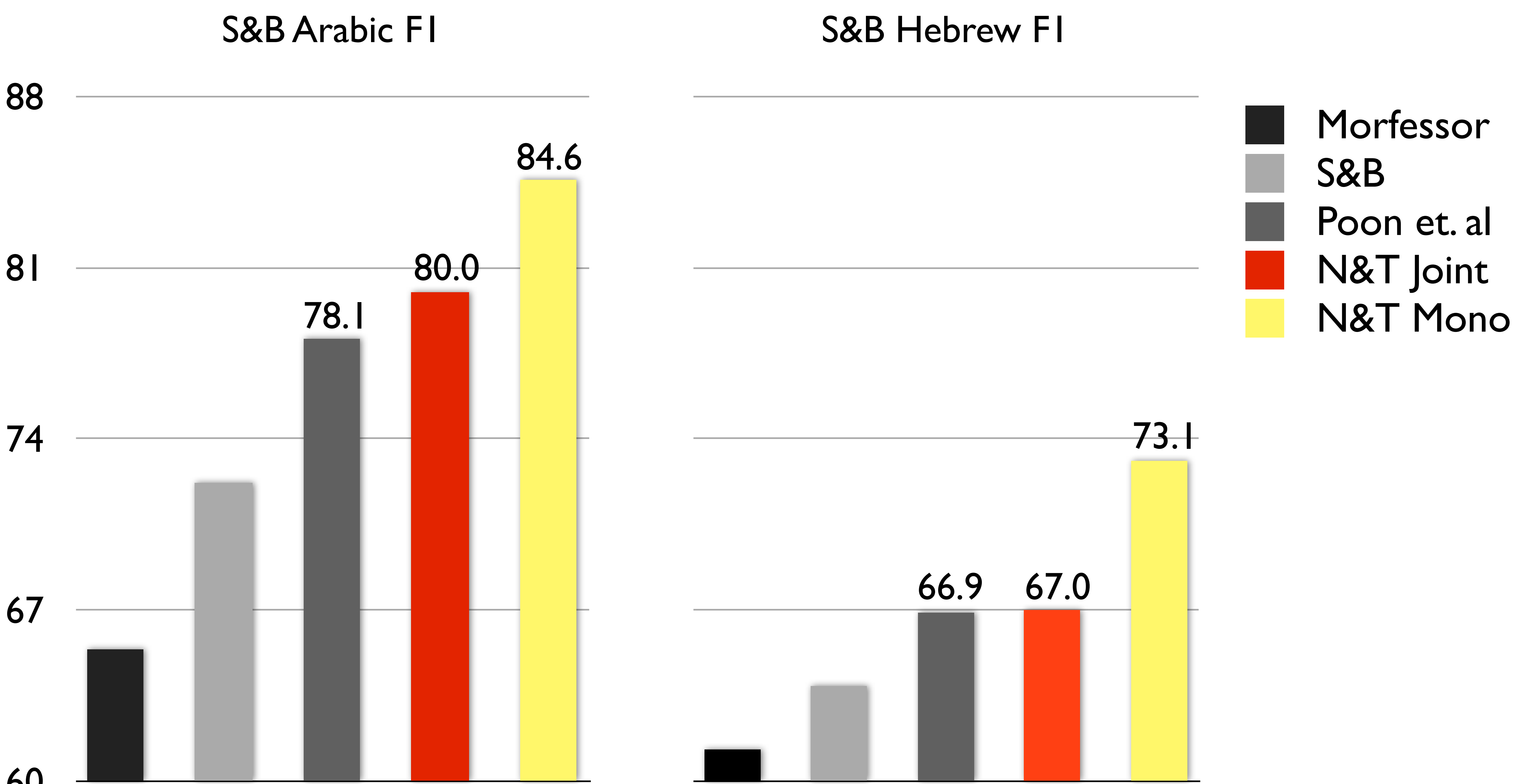
# Segmentation Experiments

- Joint Model on Hebrew & Arabic parallel phrases corpus
- Monolingual model on all 3 data sets
- Measured by F1
- Baselines:
  - Chung & Gildea (2009)
  - Morfessor
  - Snyder & Barzilay (2008)
  - Poon et. al (2009)

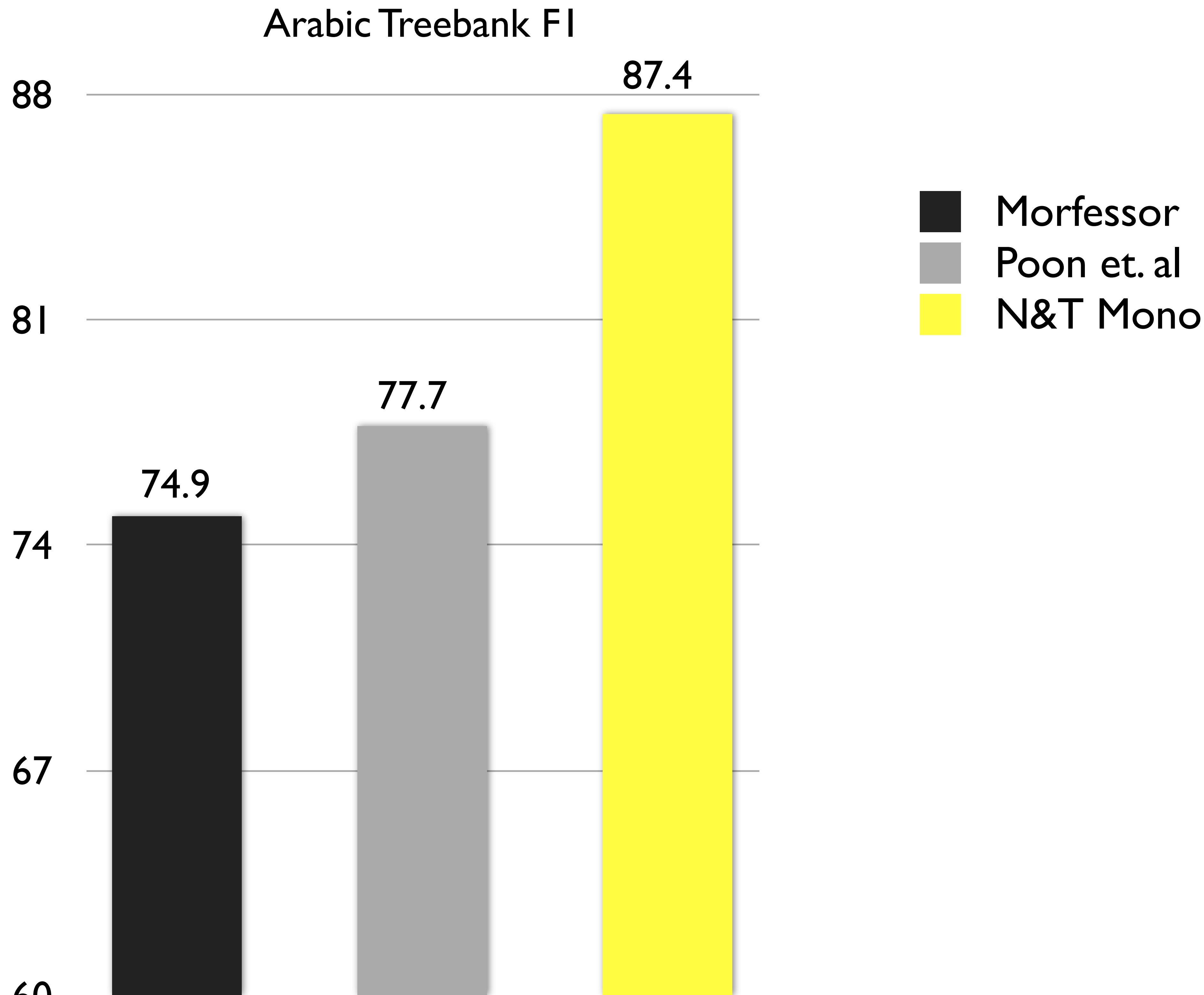
# Segmentation Results



# Segmentation Results



# Segmentation Results



# Sub-Model Results

| | Arabic | Hebrew | ATB |
|---|---|---|---|
| No Dictionary | -0.8 | -1.1 | +0.4 |
| No Boundary Modeling | -8.6 | -0.4 | -3.3 |
| No Latent Morphology | -7.4 | -13.6 | -19.5 |

(Incremental, not cumulative loss)

# Conclusions

- Contributions
  - HMM-based structure + word/morpheme aware feature-rich distortion model improves joint alignment and segmentation
  - Significant gains in morphological segmentation accuracy due to:
    - Richer Context
    - Latent morphological structure
    - Explicit modeling of word boundaries

# Future Work

- Integrate with decoding for MT
- Higher-order dependencies & morphological phenomena
- Data & human evaluations

Thank You!

благодаря ви