

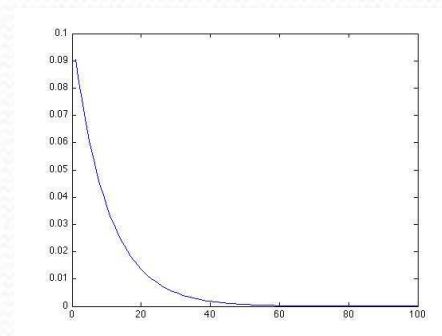
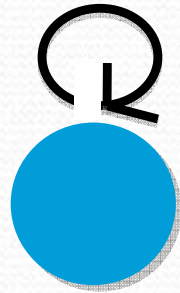
Discriminative Duration Modeling With Segmental Conditional Random Fields

Justine Kao
Geoffrey Zweig
Patrick Nguyen

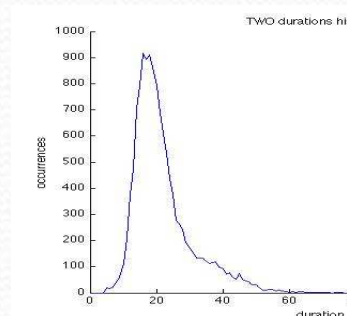
Overview

- **Motivation**
 - HMM models are inappropriate
 - The discriminative ability of duration
- **Segmental Conditional Random Fields (SCARF)**
 - Model
 - Features
- **Discriminative Duration Features in SCARF**
 - Distribution features
 - Span features
- **Results**
- **Conclusion**

- Can we do a better job at modeling word durations?
- Hidden Markov Models (HMM) use an exponential state distribution



- This doesn't match the observed distribution



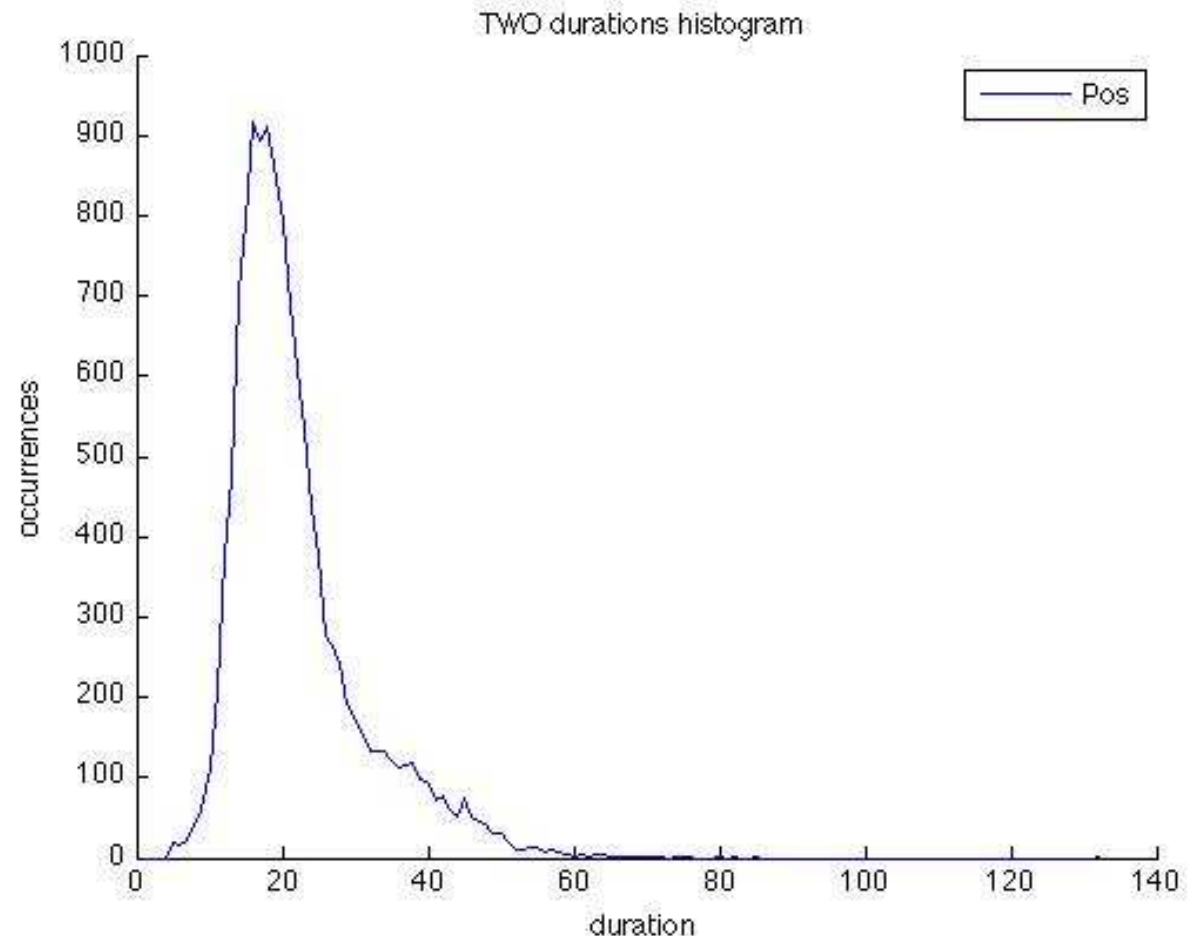
- BUT: It's possible to amend HMMs, but solutions are awkward
- BUT: Wrong doesn't mean that it will hurt recognition (“discrimination”)

Motivation

Is duration discriminative?

- For duration to be useful, the duration distributions of correct and incorrect word hypotheses need to look different

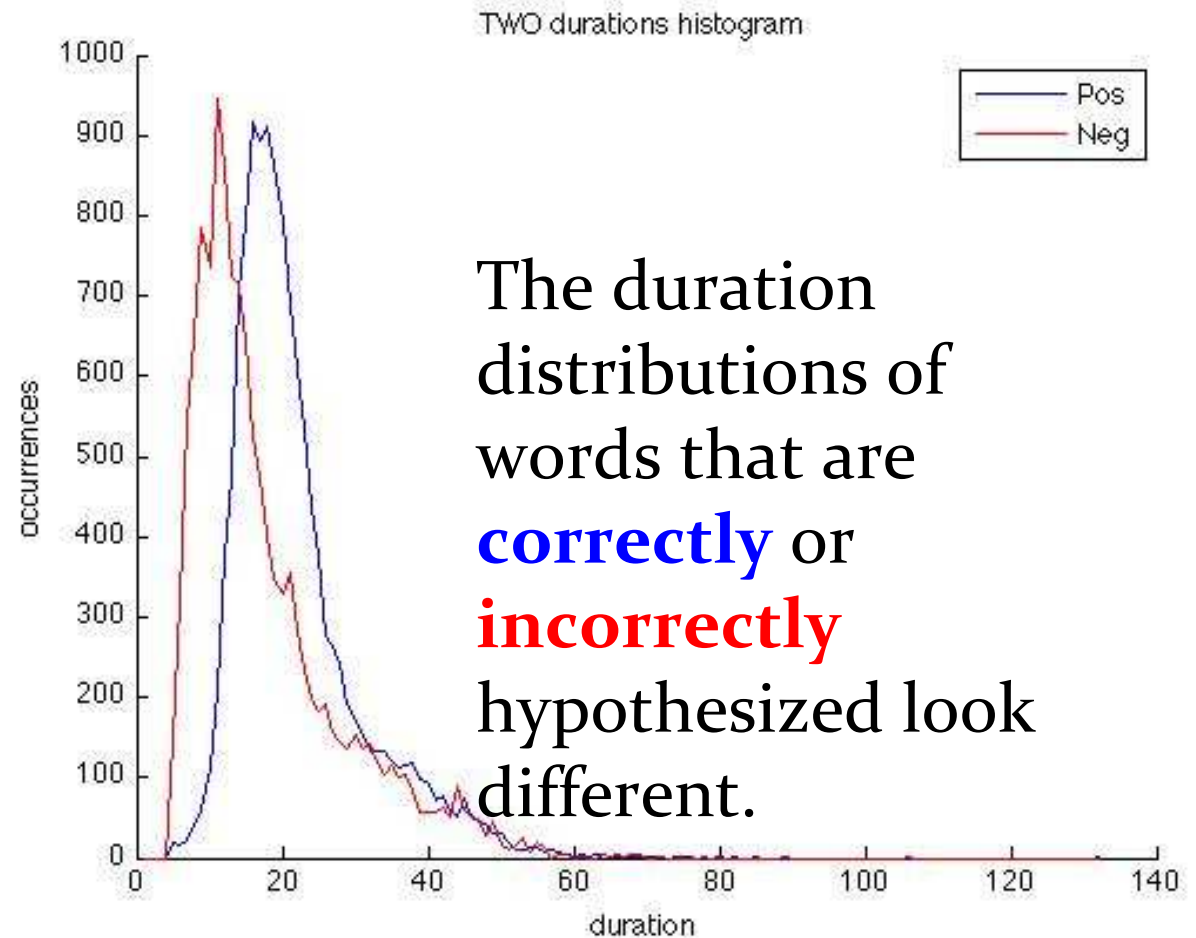
- **Are they different?**
 - Find all hypotheses of “TWO” that are **correct**
 - → positive examples
 - Plot their durations



Motivation

Is duration discriminative?

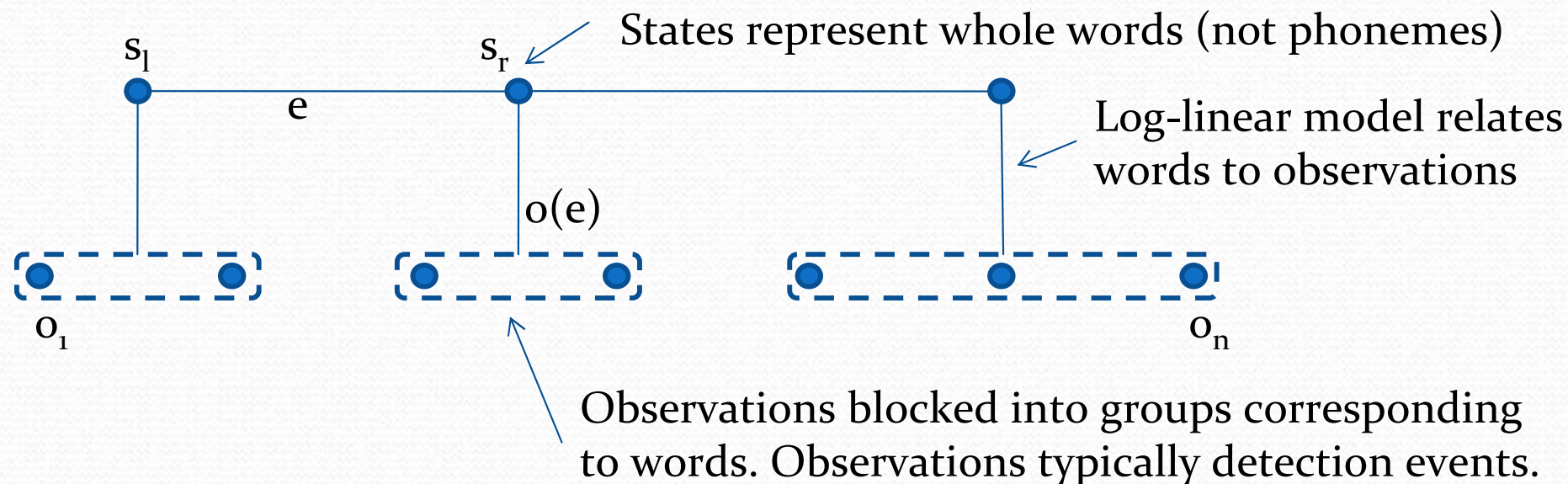
- For duration to be useful, the duration distributions of correct and incorrect word hypotheses need to look different
- **Are they different?**
 - Find all hypotheses of “TWO” that are **incorrect**
 - → negative examples
 - Plot their durations



Segmental Conditional Random Fields

How to incorporate this information?

- SCARF is well-suited for this purpose



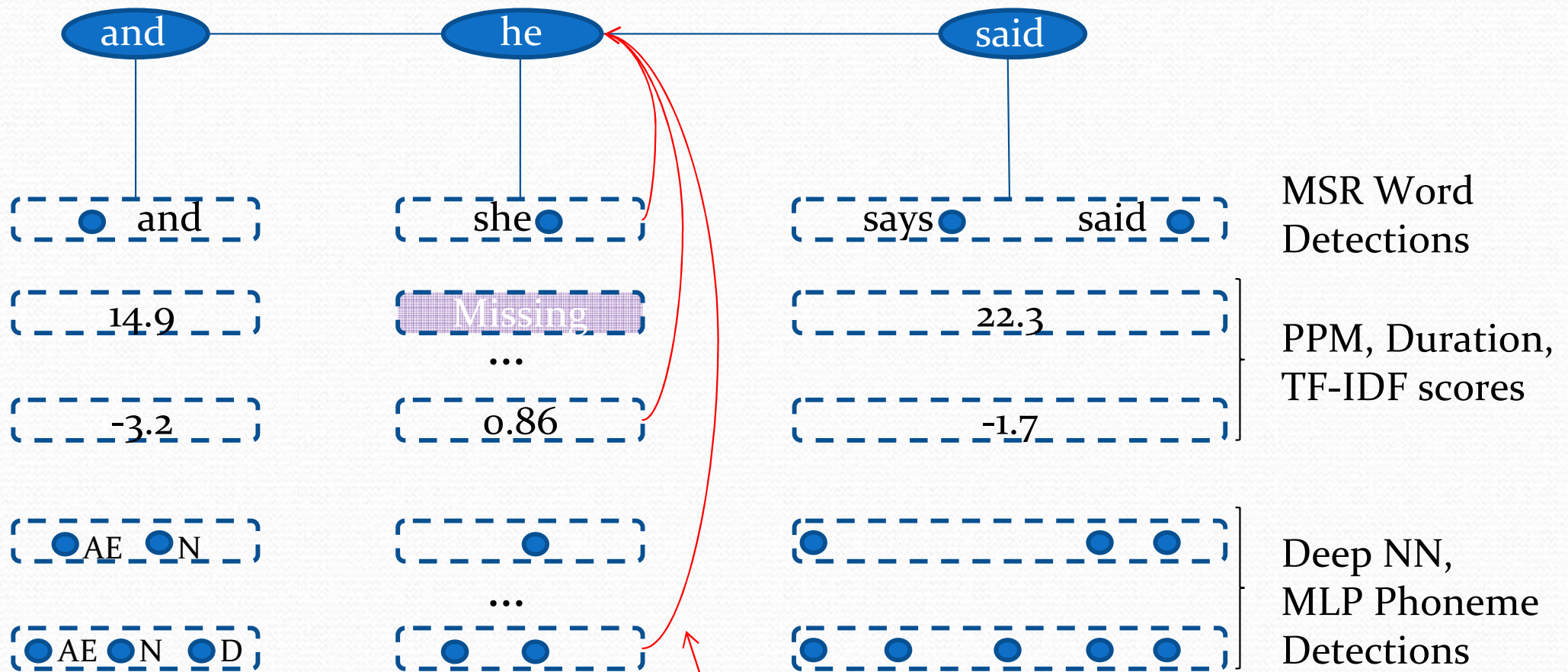
For a hypothesized word sequence s ,
we must sum over all possible segmentations q of observations

$$P(s|\mathbf{o}) = \frac{\sum_{q \text{ s.t. } |q|=|s|} \exp(\sum_{e \in q, k} \lambda_k f_k(s_l^e, s_r^e, o(e)))}{\sum_{s'} \sum_{q \text{ s.t. } |q|=|s'|} \exp(\sum_{e \in q, k} \lambda_k f_k(s_l'^e, s_r'^e, o(e)))}$$

Segmental Conditional Random Fields

An integrating framework

- See JHU workshop poster this afternoon

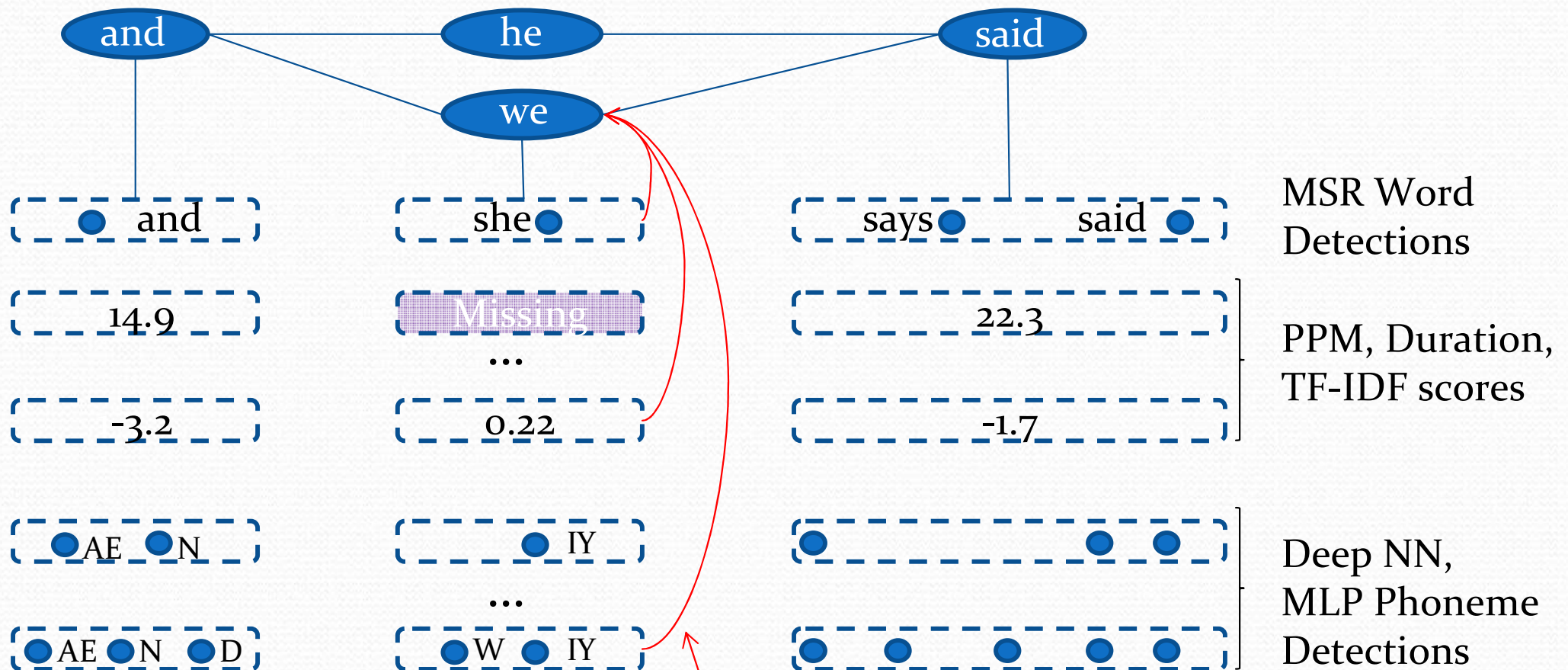


Features measure consistency
between observations & hypothesis

Segmental Conditional Random Fields

An integrating framework

- See JHU workshop poster this afternoon



Segmental Conditional Random Fields

Penalizing paths with word scores

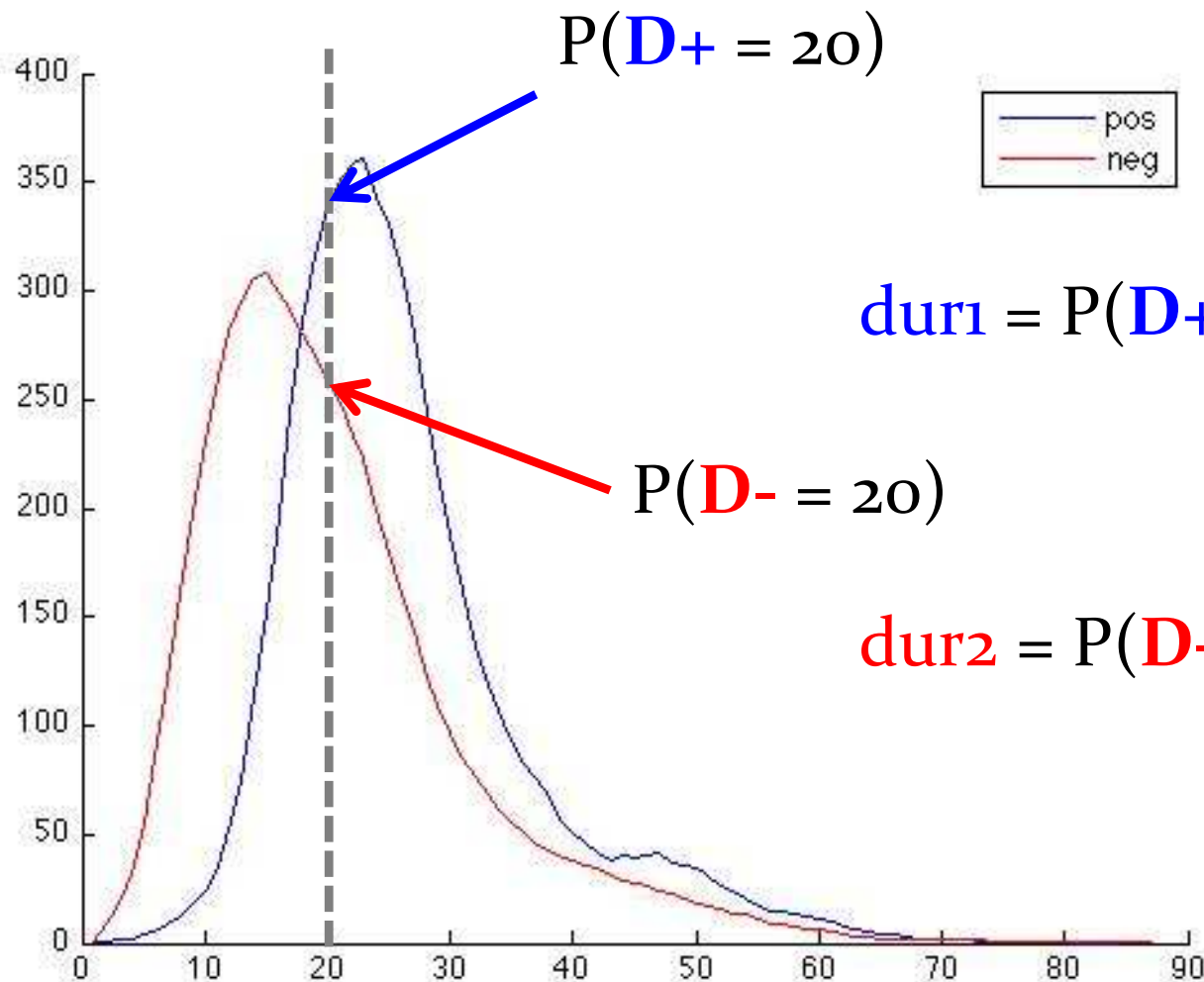
- The language model tries to insert many short frequent words, making them shorter to “fit”
- Penalize them based on their short duration
- SCARF gives us the ability to add a penalty to hypotheses



Duration as feature

Probability density function scores

- **Task:** given the duration of a word hypothesis, capture the likelihood of it being in the correct or incorrect distribution
 - Suppose a word hypothesis “TWO” is 20 frames long



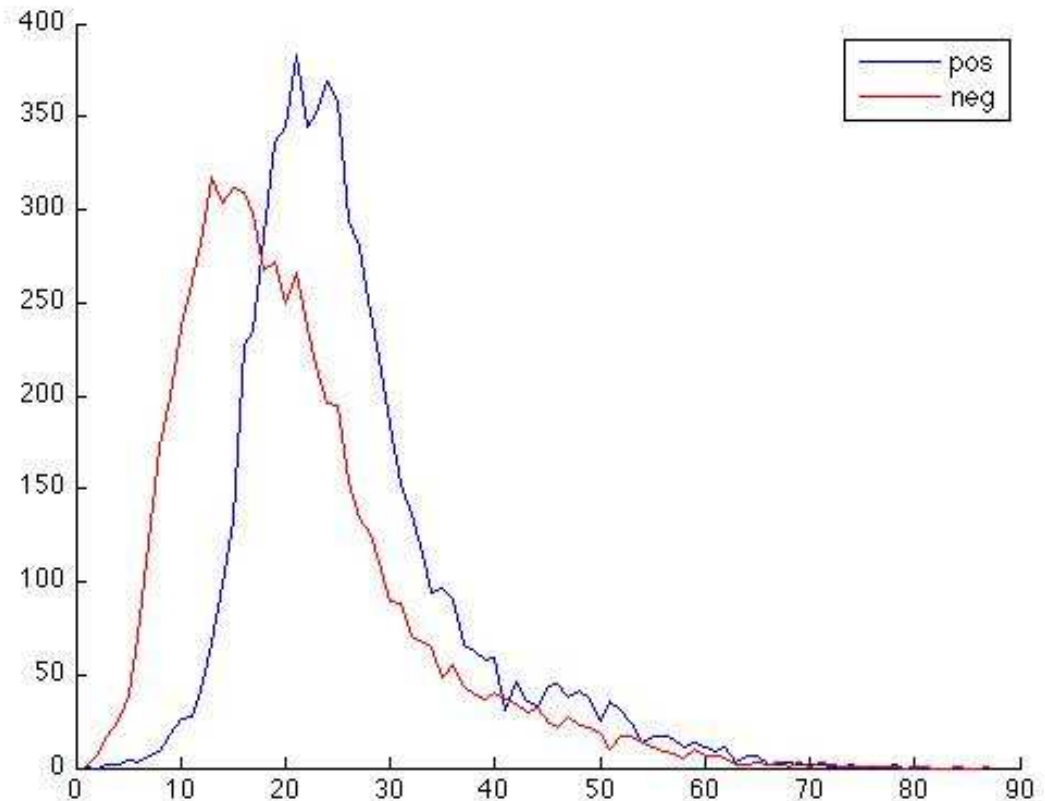
$$\text{dur1} = P(\mathbf{D}_+ = 20) = \frac{\text{num}(\mathbf{D}_+ = 20)}{\text{num}(\text{"TWO"} +)}$$

$$\text{dur2} = P(\mathbf{D}_- = 20) = \frac{\text{num}(\mathbf{D}_- = 20)}{\text{num}(\text{"TWO"} -)}$$

Duration as a feature

Focus on the top 100 words

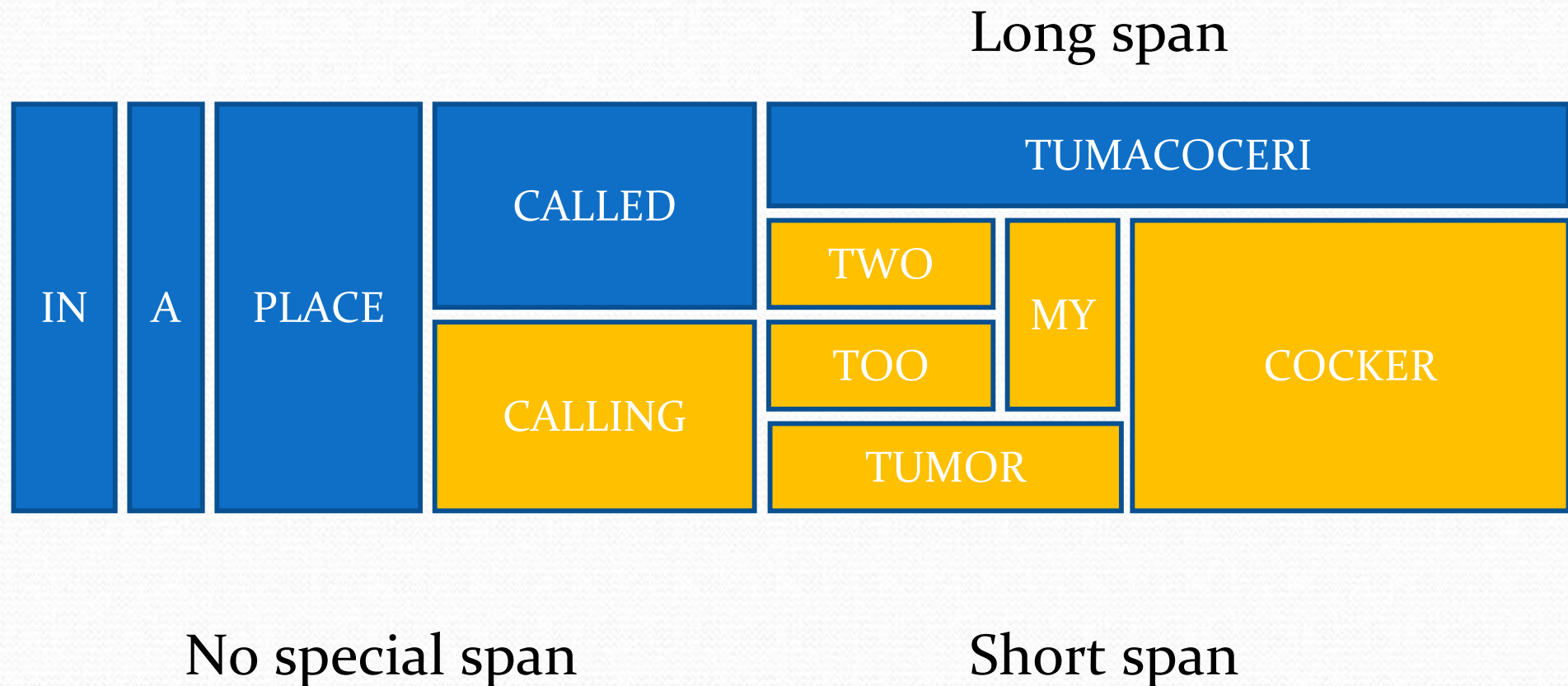
- Focus on the top 100 most frequent words seen in the training transcripts
- **Large portion of data:** The top 100 most frequent words account for 47.5% of all word occurrences in the training set transcript
- **Large portion of important data:** The top 100 most frequent words account for 48.58% of all errors in the test set
 - Function words, shorter



Duration features

Span features

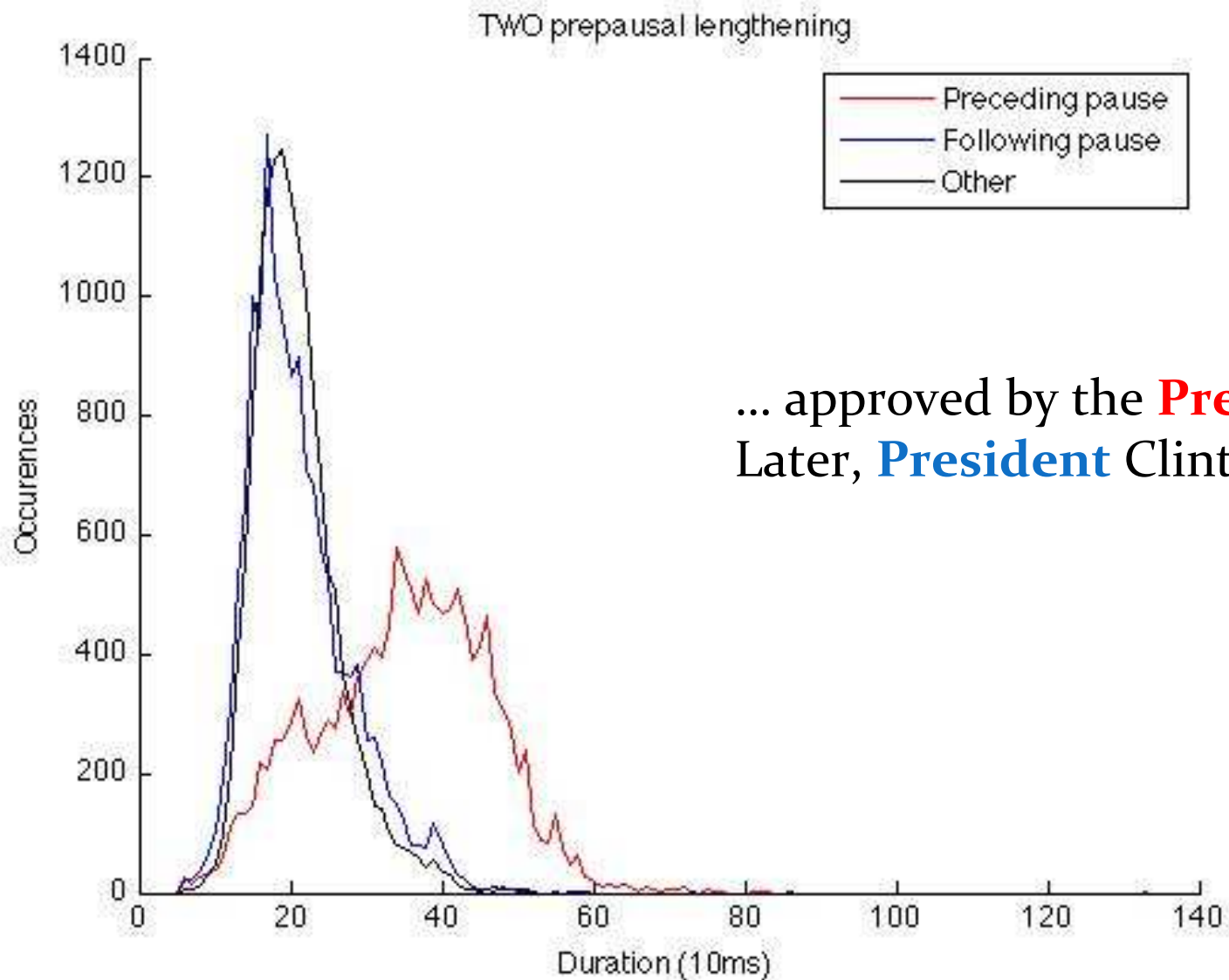
- Duration can be treated separately in the case of an expected substitution or insertion of tiny fragments
- This gives us the ability to boost “Tumacoceri” vs other instances



Duration features

Pre-pausal lengthening

- Words are stretched before a pause (e.g. Gadde, 2000)



... approved by the **President**. *pause*
Later, **President** Clinton said ...

Experimental results

Over a state-of-the-art system

- State-of-the-art IBM baseline (Attila)
- Combined with MSR system
- Small gain, as good as any other individual feature we tried at the JHU workshop

Broadcast News	Devo4f
Baseline	16.3% WER
+MSR word detectors	15.3
+Duration	15.2
& Pause context	15.1
& spanning	15.0

Conclusion

Word durations help discrimination

- Derived insights about discriminative ability of duration
 - Duration of correctly recognized words differs from duration of incorrectly recognized words, typically longer
 - Language model encourages insertion of small function words
- Formulated insights as features in Segmental Conditional Random Fields
 - Penalizes spurious word hypotheses with duration scores
 - SCARF makes it easy to combine in current system
- Improvement over a state-of-the-art Broadcast News setup