Contextual Information Improves OOV Detection in Speech

Carolina Parada

Center for Language and Speech Processing
The Johns Hopkins University
Joint work with Frederick Jelinek, Mark Dredze, and Denis Filimonov

PIRE Meeting December 12, 2009

Outline

- Motivation
- Hybrid system for OOV detection (baseline)
- Overview: Conditional Random Fields vs Maximum Entropy
- Experimental Setup and Evaluation
- Improvements over baseline system
  - Generalize to sequence labeling problem
  - Incorporate local lexical information
  - Incorporate global information
- Conclusions and future work

Outline

- Motivation
- Hybrid system for OOV detection (baseline)
- Overview: Conditional Random Fields vs Maximum Entropy
- Experimental Setup and Evaluation
- Improvements over baseline system
  - Generalize to sequence labeling problem
  - Incorporate local lexical information
  - Incorporate global information
- Conclusions and future work

Automatic Speech Recognition (ASR)

- Find most likely word sequence uttered by a speaker given the acoustics.

$$\hat{W} = \arg\max_{W} P(W|A)$$

Automatic Speech Recognition (ASR)

- Find most likely word sequence uttered by a speaker given the acoustics.

$$\hat{W} = \arg\max_W P(W|A) = \arg\max_W P(W)P(A|W)$$

Automatic Speech Recognition (ASR)

- Find most likely word sequence uttered by a speaker given the acoustics.

language model

acoustic model

$$\hat{W} = \arg\max_W P(W|A) = \arg\max_W P(W)P(A|W)$$

search

Automatic Speech Recognition (ASR)

- Find most likely word sequence uttered by a speaker given the acoustics.

language model

acoustic model

$$\hat{W} = \arg\max_W P(W|A) = \arg\max_W P(W)P(A|W)$$

search

- Assumption: words belongs to a finite vocabulary.

Automatic Speech Recognition (ASR)

- Find most likely word sequence uttered by a speaker given the acoustics.

language model

acoustic model

$$\hat{W} = \arg\max_W P(W|A) = \arg\max_W P(W)P(A|W)$$

search

- Assumption: words belongs to a finite vocabulary.
- Even with a large vocabulary, ASR systems will encounter words not seen during training: Out-Of-Vocabulary (OOV) words.

Why are OOVs important?

- OOVs are an important source of error in ASR systems:
  - They can never be recognized by the basic system, even if repeated.
  - They contribute to recognition errors in surrounding words.
    - Causes error-propagation for downstream applications.
  - OOVs are often information-rich nouns: named-entities, foreign words.

Why are OOVs important?

- OOVs are an important source of error in ASR systems:
  - They can never be recognized by the basic system, even if repeated.
  - They contribute to recognition errors in surrounding words.
    - Causes error-propagation for downstream applications.
    - OOVs are often information-rich nouns: named-entities, foreign words.

- Why not just increase the vocabulary?
  - There will always be words the system has never encountered.
  - Increasing vocabulary size without limit can lead to higher word-error-rates (WER).

Why are OOVs important?

- OOVs are an important source of error in ASR systems:
  - They can never be recognized by the basic system, even if repeated.
  - They contribute to recognition errors in surrounding words.
    - Causes error-propagation for downstream applications.
    - OOVs are often information-rich nouns: named-entities, foreign words.

- Why not just increase the vocabulary?
  - There will always be words the system has never encountered.
  - Increasing vocabulary size without limit can lead to higher word-error-rates (WER).

Detect OOV regions in the output of the ASR system

What is the task?

- Task: detect OOVs in the output of an ASR system.

truth: disappearance of a north dakota college student

What is the task?

- Task: detect OOVs in the output of an ASR system.

truth: disappearance of a north dakota college student

output: disappearance of a north to coat a college student

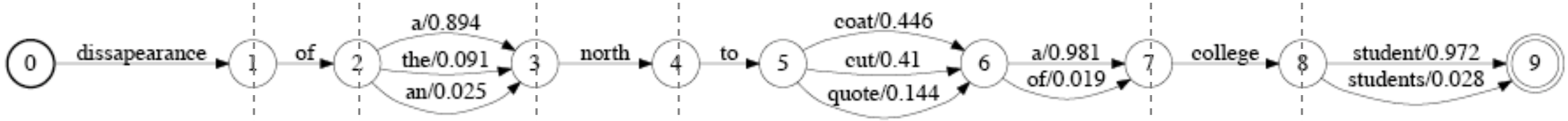
What is the task?

- Task: detect OOVs in the output of an ASR system.
- Confusion network (CN): compact representation of ASR hypothesis.*

truth: disappearance | of | a | north | dakota | college | student

output: disappearance | of | a | north | to | coat | a | college | student

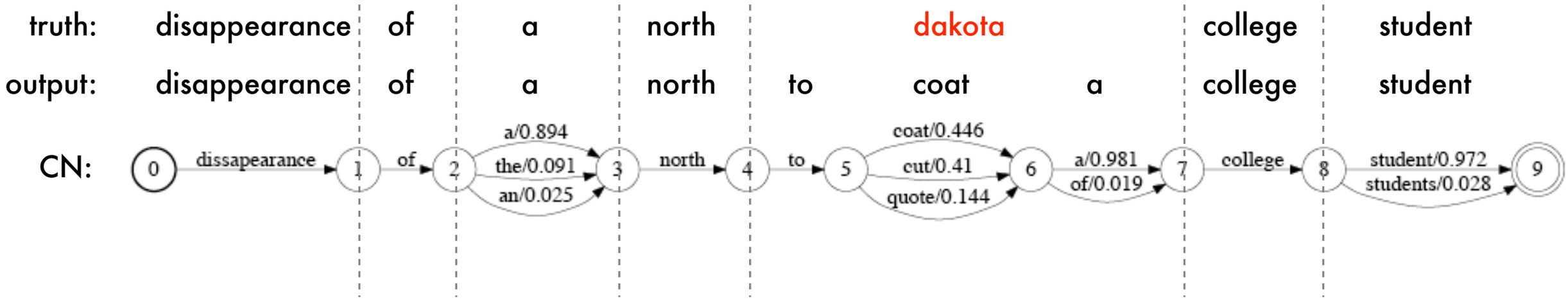
CN:



*(Mangu et al 1999)

What is the task?

- Task: detect OOVs in the output of an ASR system.
- Confusion network (CN): compact representation of ASR hypothesis.*
- Given confusion network, label each region in the network as IV/OOV.



*(Mangu et al 1999)

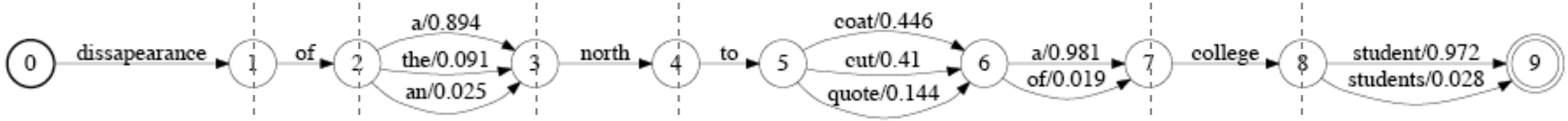
What is the task?

- Task: detect OOVs in the output of an ASR system.
- Confusion network (CN): compact representation of ASR hypothesis.*
- Given confusion network, label each region in the network as IV/OOV.

truth: ~~disappearance~~ | ~~of~~ | ~~a~~ | ~~north~~ | ~~dakota~~ | ~~college~~ | ~~student~~

output: disappearance | of | a | north | to | coat | a | college | student

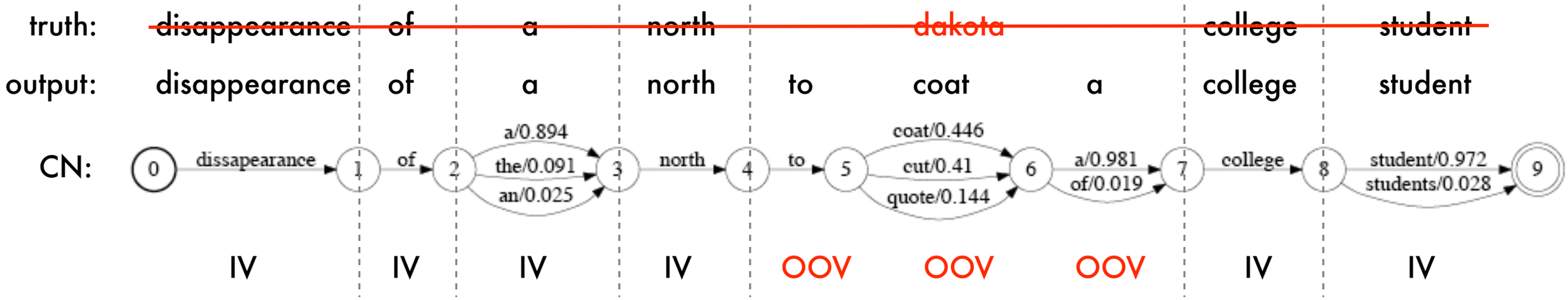
CN:



*(Mangu et al 1999)

What is the task?

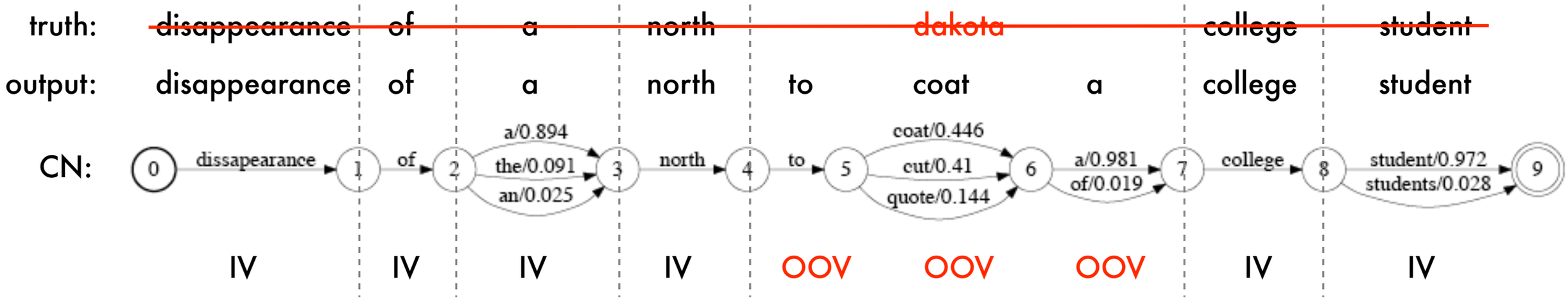
- Task: detect OOVs in the output of an ASR system.
- Confusion network (CN): compact representation of ASR hypothesis.*
- Given confusion network, label each region in the network as IV/OOV.



*(Mangu et al 1999)

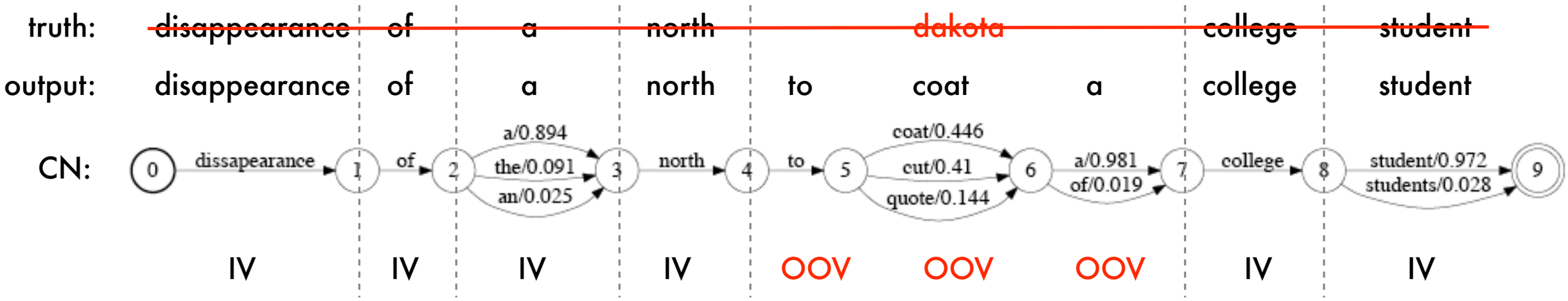
What is the task?

- Task: detect OOVs in the output of an ASR system.
- Confusion network (CN): compact representation of ASR hypothesis.*
- Given confusion network, label each region in the network as IV/OOV.



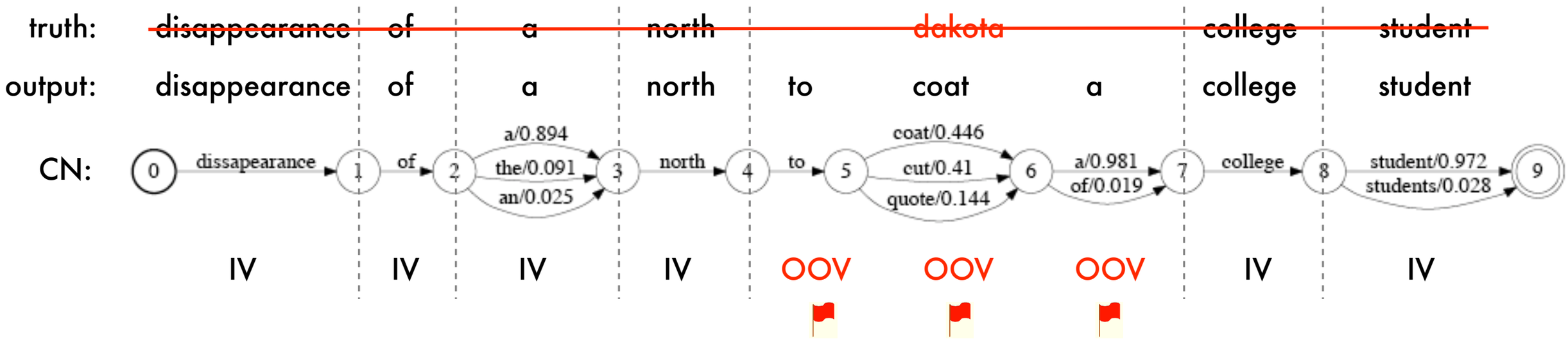
*(Mangu et al 1999)

Why is it useful to detect OOVs?



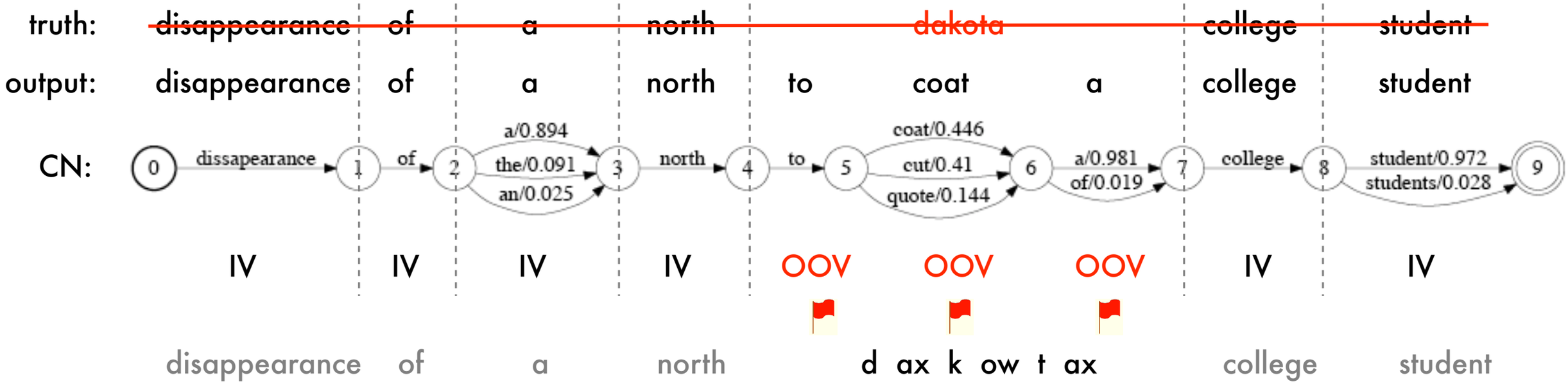
*(Mangu et al 1999)

Why is it useful to detect OOVs?



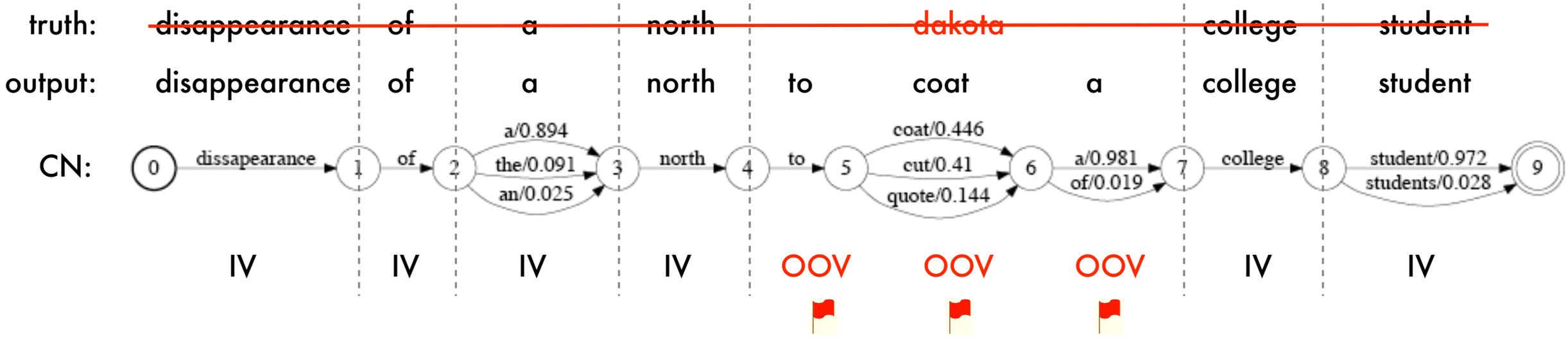
- Flag for annotation, or inclusion in the vocabulary.

Why is it useful to detect OOVs?



- Flag for annotation, or inclusion in the vocabulary.
- Transcribe OOV segments using a phonetic recognizer: open vocabulary ASR .

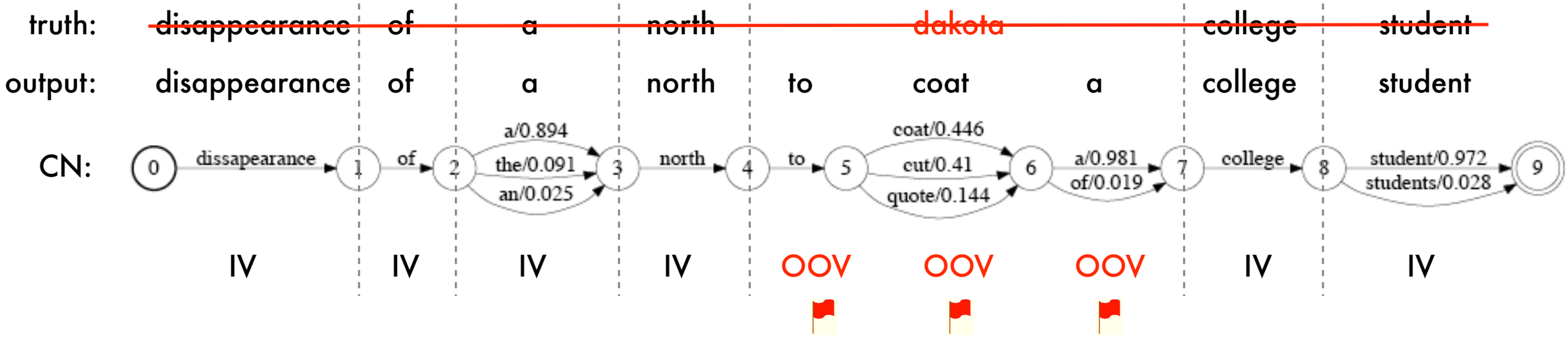
Why is it useful to detect OOVs?



disappearance	of	a	north	d ax k ow t ax	college	student
disappearance	of	a	north	{OOV,NAME,ORG, LOC}	college	student

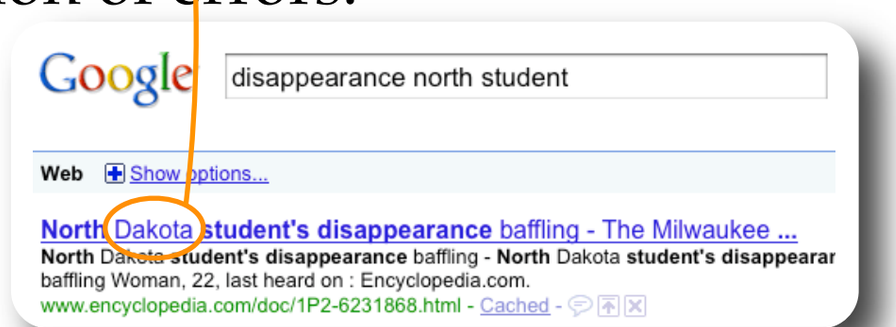
- Flag for annotation, or inclusion in the vocabulary.
- Transcribe OOV segments using a phonetic recognizer: open vocabulary ASR.
- Replace OOV segment as {OOV} to avoid propagation of errors.

Why is it useful to detect OOVs?

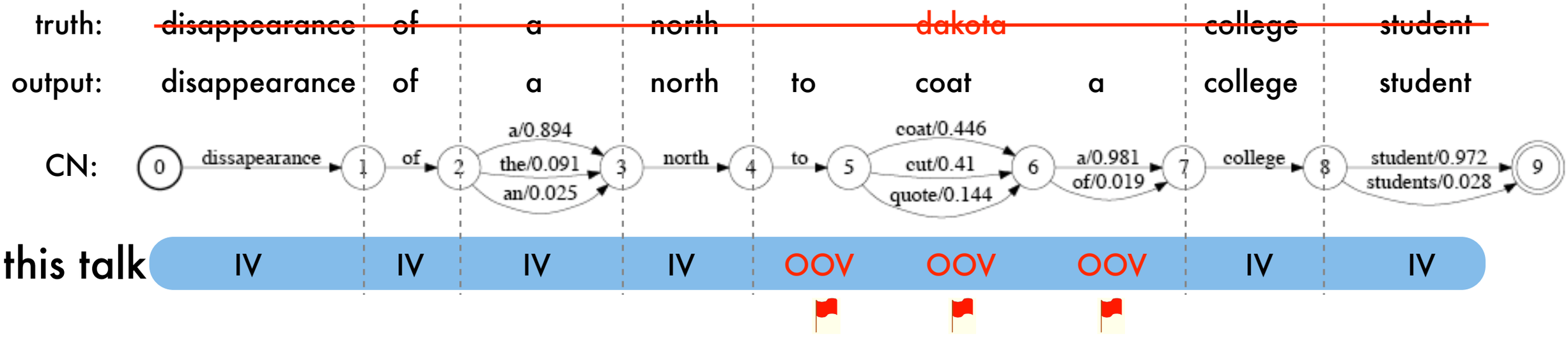


disappearance	of	a	north	d ax k ow t ax	college	student
disappearance	of	a	north	{OOV,NAME,ORG, LOC}	college	student
disappearance	of	a	north	Dakota	college	student

- Flag for annotation, or inclusion in the vocabulary.
- Transcribe OOV segments using a phone recognizer: open vocabulary ASR.
- Replace OOV segment as {OOV} to avoid propagation of errors.
- Use context + web + acoustics in OOV-region to recover OOVs.

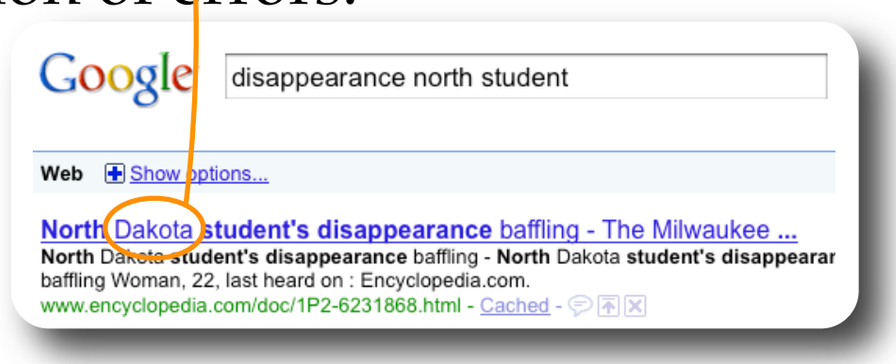


Why is it useful to detect OOVs?



disappearance	of	a	north	d ax k ow t ax	college	student
disappearance	of	a	north	{OOV,NAME,ORG, LOC}	college	student
disappearance	of	a	north	Dakota	college	student

- Flag for annotation, or inclusion in the vocabulary.
- Transcribe OOV segments using a phone recognizer: open vocabulary ASR.
- Replace OOV segment as {OOV} to avoid propagation of errors.
- Use context + web + acoustics in OOV-region to recover OOVs.



Outline

- Motivation
- Hybrid system for OOV detection (baseline)
- Overview: Conditional Random Fields vs Maximum Entropy
- Experimental Setup and Evaluation
- Improvements over baseline system
  - Generalize to sequence labeling problem
  - Incorporate local lexical information
  - Incorporate global information
- Conclusions and future work

Previous Work: Hybrid System

- Models OOVs by combining word and sub-word units (*fragments*) in vocabulary (Rastrow et. al. 2009).
  - Fragments are variable-length phone sequences (data-driven).
  - Fragments are used to represent OOVs in the Language Model text.

Previous Work: Hybrid System

- Models OOVs by combining word and sub-word units (*fragments*) in vocabulary (Rastrow et. al. 2009).
  - Fragments are variable-length phone sequences (data-driven).
  - Fragments are used to represent OOVs in the Language Model text.

word: n. b. c.’s jim miklaszewski has the latest on ...
hybrid: n. b. c.’s jim m_ik_k l_ax sh_eh_f s_k iy has the latest on ...

Previous Work: Hybrid System

- Models OOVs by combining word and sub-word units (*fragments*) in vocabulary (Rastrow et. al. 2009).
  - Fragments are variable-length phone sequences (data-driven).
  - Fragments are used to represent OOVs in the Language Model text.

word: n. b. c.'s jim miklaszewski has the latest on ...
hybrid: n. b. c.'s jim m_ik_k l_ax sh_eh_f s_k iy has the latest on ...

- Output Confusion Networks include words and fragments.

Previous Work: Hybrid System

- Models OOVs by combining word and sub-word units (*fragments*) in vocabulary (Rastrow et. al. 2009).
  - Fragments are variable-length phone sequences (data-driven).
  - Fragments are used to represent OOVs in the Language Model text.

word: n. b. c.'s jim miklaszewski has the latest on ...
hybrid: n. b. c.'s jim m_ik_k l_ax sh_eh_f s_k iy has the latest on ...

- Output Confusion Networks include words and fragments.
- To detect OOVs: combines probability of fragments and other confidence measures.

Previous Work: Hybrid System

- Models OOVs by combining word and sub-word units (*fragments*) in vocabulary (Rastrow et. al. 2009).
  - Fragments are variable-length phone sequences (data-driven).
  - Fragments are used to represent OOVs in the Language Model text.

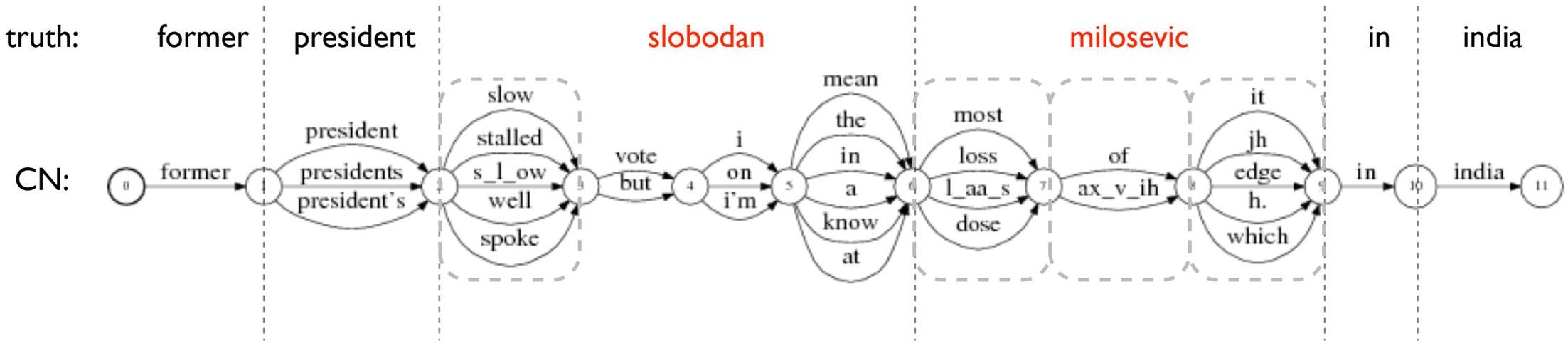
word: n. b. c.'s jim miklaszewski has the latest on ...
hybrid: n. b. c.'s jim m_ik_k l_ax sh_eh_f s_k iy has the latest on ...

- Output Confusion Networks include words and fragments.
- To detect OOVs: combines probability of fragments and other confidence measures.
- Achieves state-of-the-art performance using only a few features.

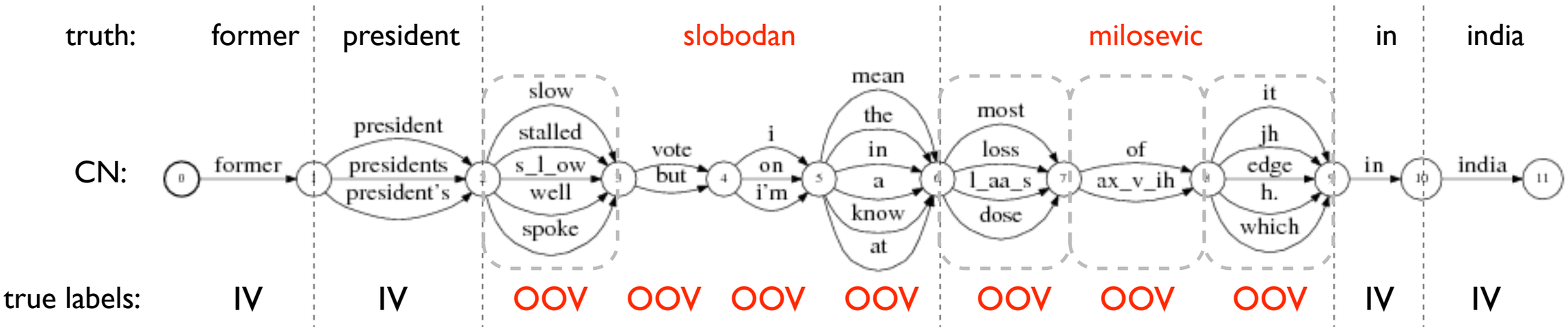
Previous Work: Hybrid System

truth: former president slobodan milosevic in india

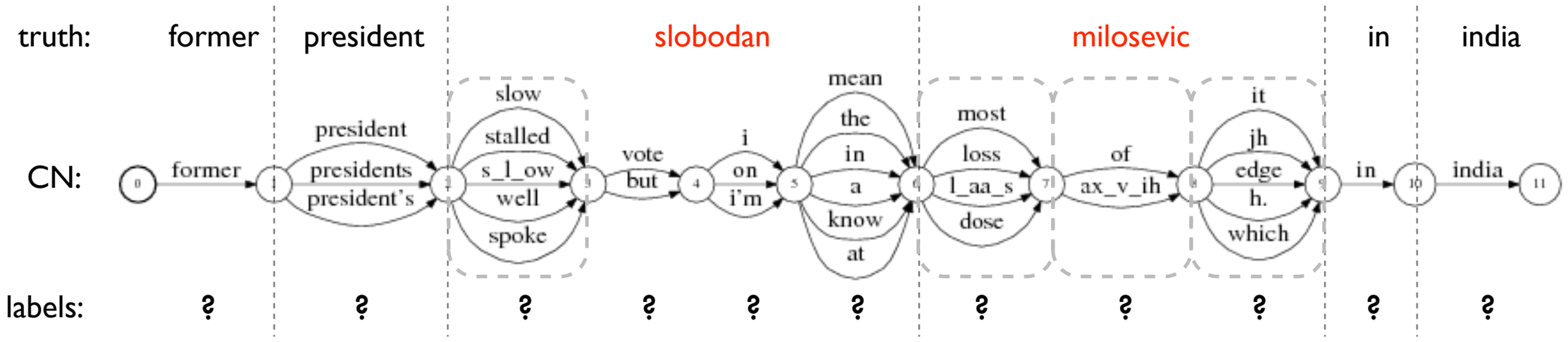
Previous Work: Hybrid System



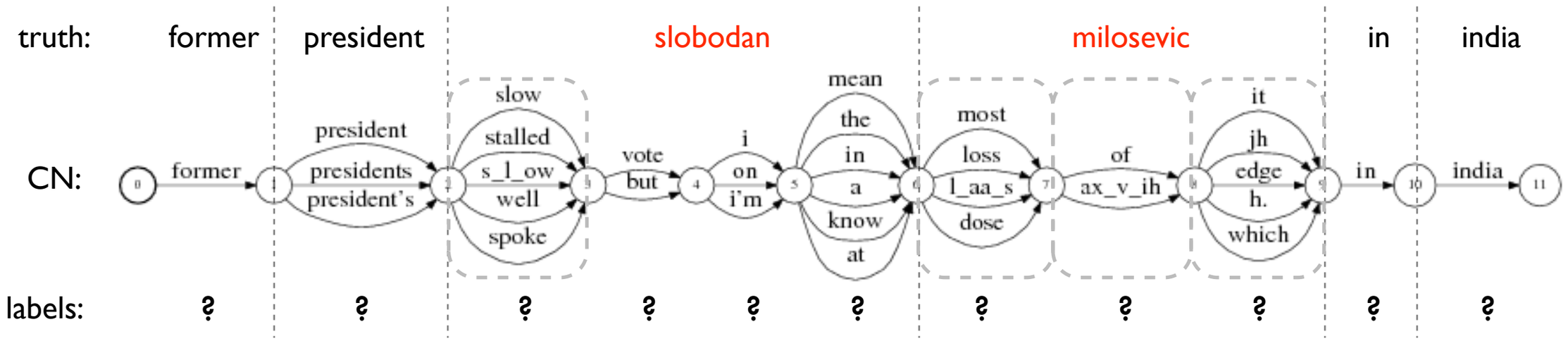
Previous Work: Hybrid System



Previous Work: Hybrid System

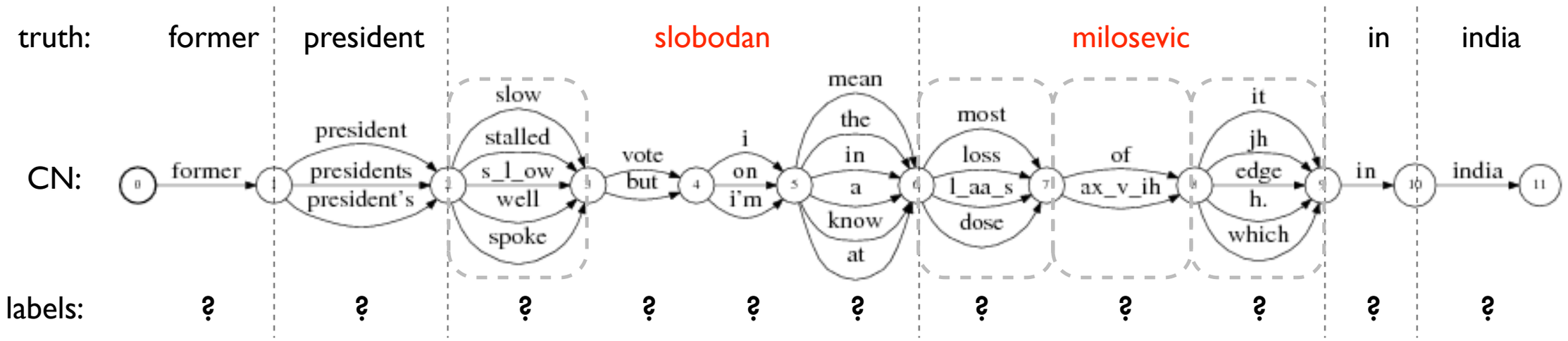


Previous Work: Hybrid System



- Treat OOV detection as binary-classification task on each confusion region/bin.

Previous Work: Hybrid System



- Treat OOV detection as binary-classification task on each confusion region/bin.
- Features used:

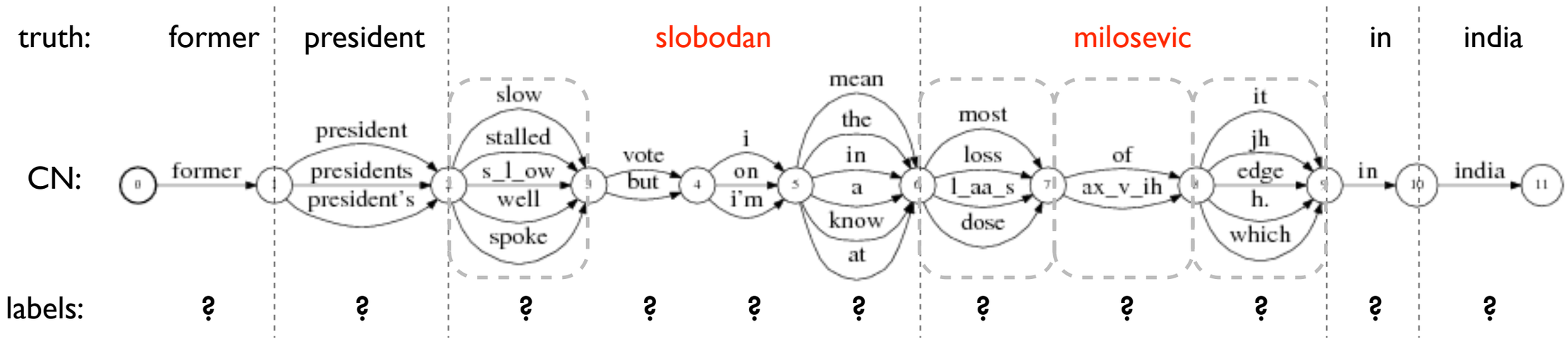
fragment

confusion bin

$$\text{Fragment-Posterior} = \sum_{f \in \{t_j\}} p(f|t_j)$$

$$\text{Word-Entropy} = -\sum_{w \in t_j} p(w|t_j) \log p(w|t_j)$$

Previous Work: Hybrid System



- Treat OOV detection as binary-classification task on each confusion region/bin.
- Features used:

fragment

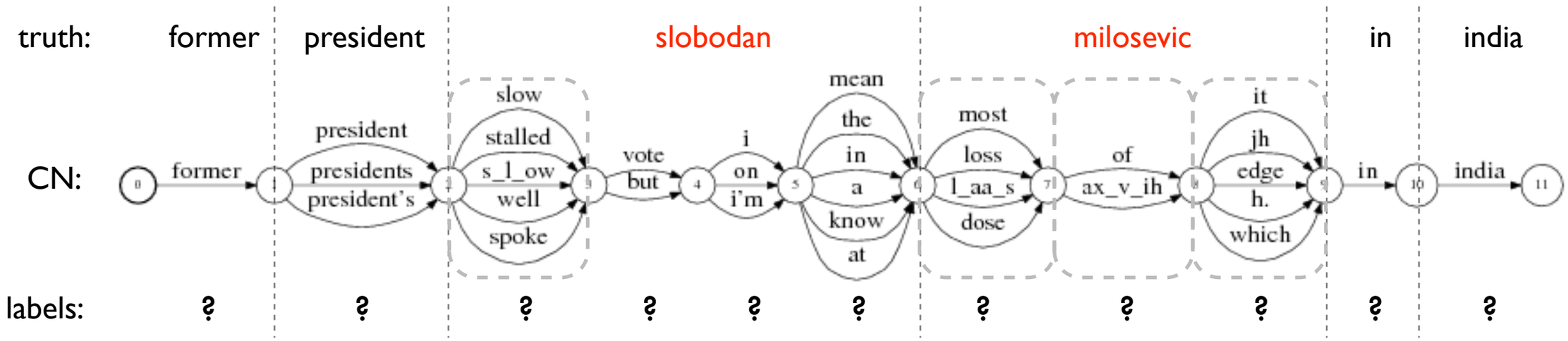
confusion bin

$$\text{Fragment-Posterior} = \sum_{f \in \{t_j\}} p(f|t_j)$$

$$\text{Word-Entropy} = -\sum_{w \in t_j} p(w|t_j) \log p(w|t_j)$$

- Combines features using Maximum Entropy classifier.

Previous Work: Hybrid System



- Treat OOV detection as binary-classification task on each confusion region/bin.

- Features used:

fragment

confusion bin

$$\text{Fragment-Posterior} = \sum_{f \in \{t_j\}} p(f|t_j)$$

$$\text{Word-Entropy} = -\sum_{w \in t_j} p(w|t_j) \log p(w|t_j)$$

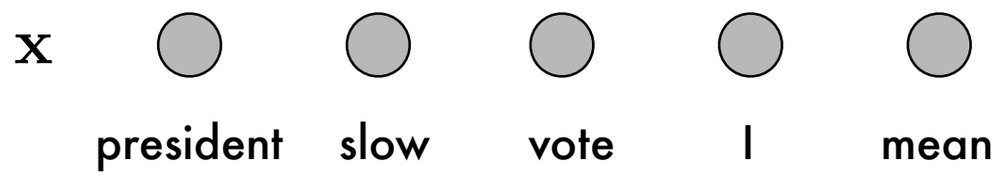
- Combines features using Maximum Entropy classifier.

- This approach classifies each region independently using local information.

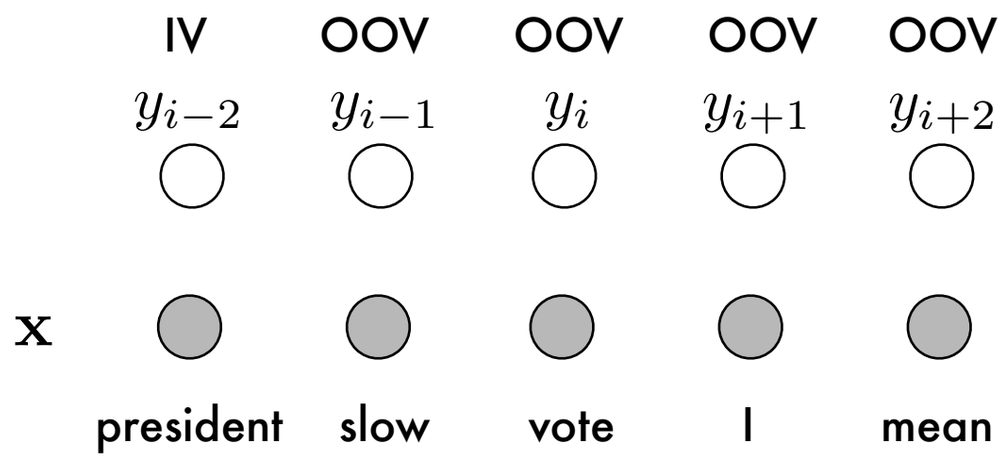
Maximum Entropy Classifier

president slow vote I mean

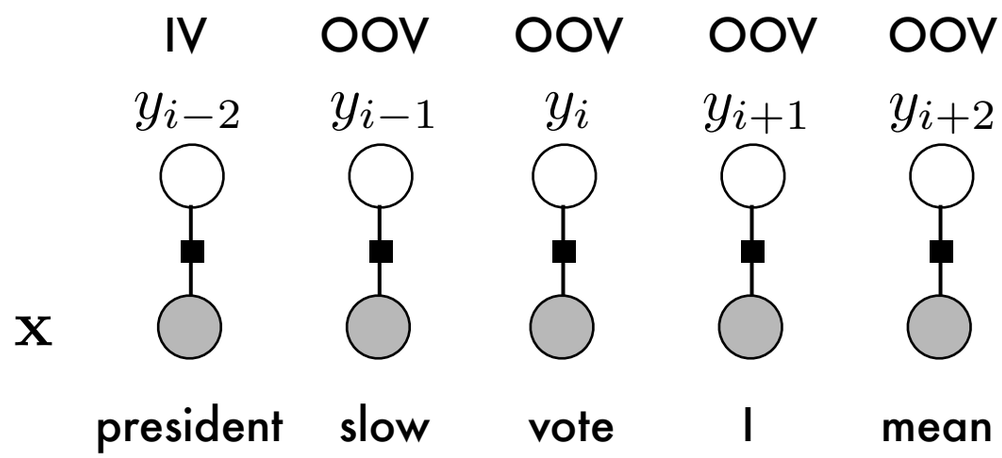
Maximum Entropy Classifier



Maximum Entropy Classifier

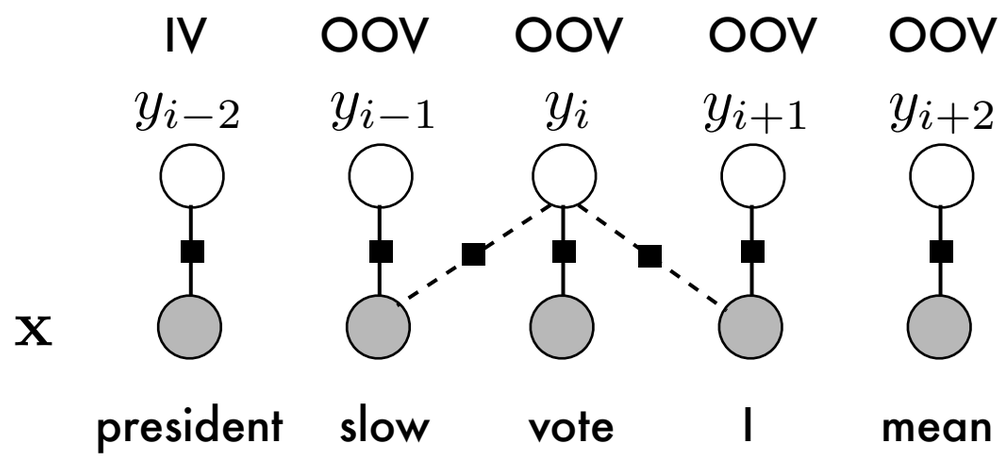


Maximum Entropy Classifier



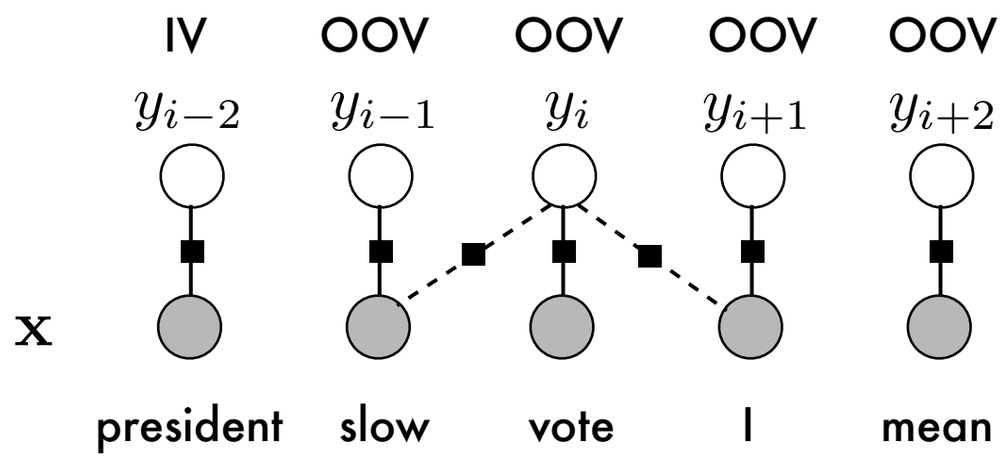
- Discriminative classifier: predicts hidden labels given observed sequence.

Maximum Entropy Classifier



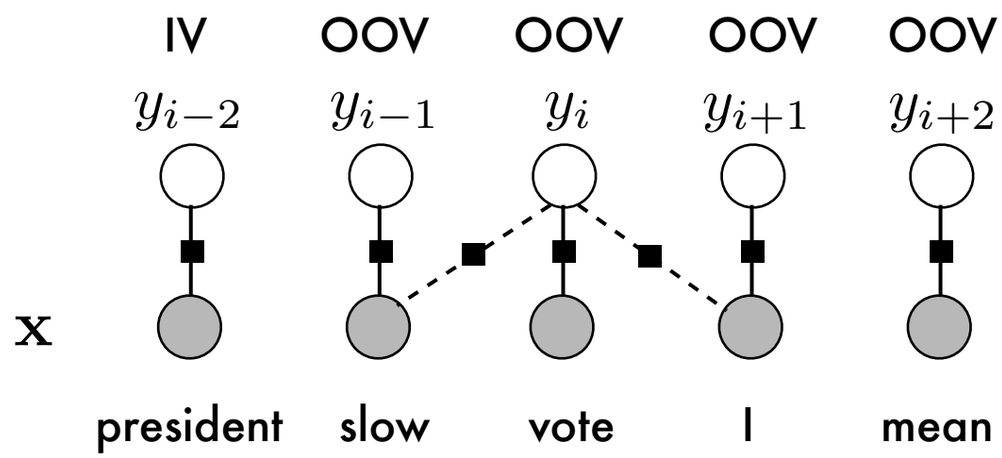
- Discriminative classifier: predicts hidden labels given observed sequence.

Maximum Entropy Classifier



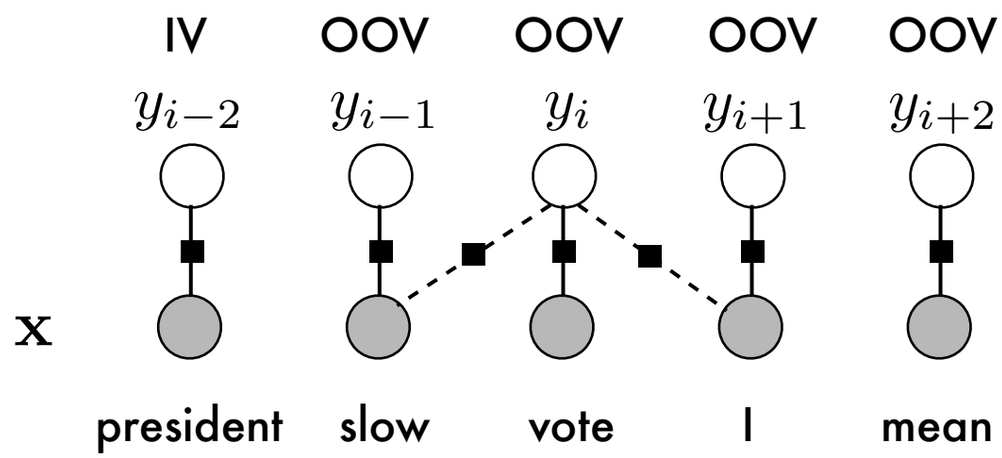
- Discriminative classifier: predicts hidden labels given observed sequence.
- Assigns a label to each region independently.

Maximum Entropy Classifier



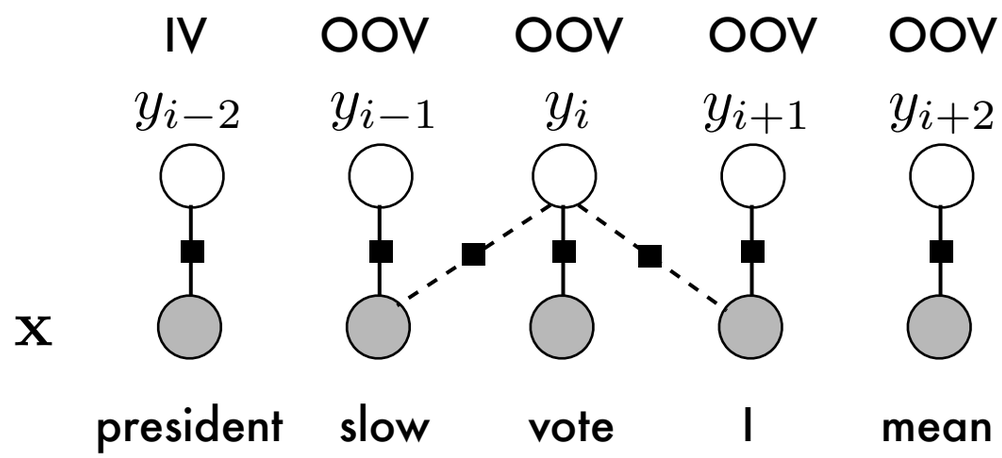
- Discriminative classifier: predicts hidden labels given observed sequence.
- Assigns a label to each region independently.
- Often used for OOV and error detection (Burget et al 08, Rastrow et al. 09, White et al 07, Hazen and Bassi 2001)

Maximum Entropy Classifier



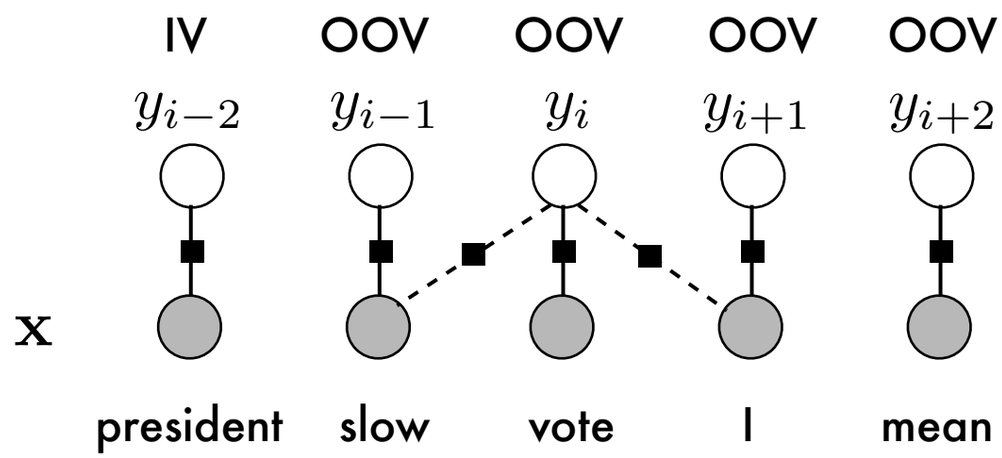
- Discriminative classifier: predicts hidden labels given observed sequence.
- Assigns a label to each region independently.
- Often used for OOV and error detection (Burget et al 08, Rastrow et al. 09, White et al 07, Hazen and Bassi 2001)
- Problems:

Maximum Entropy Classifier



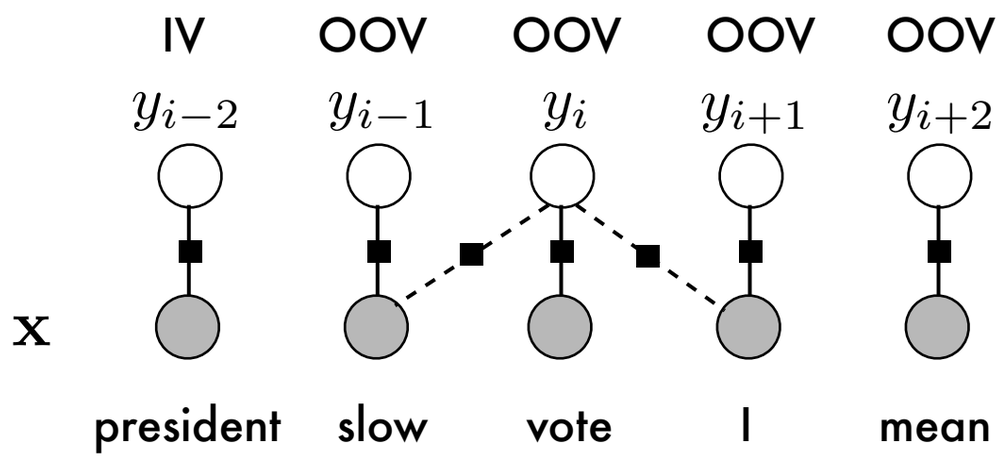
- Discriminative classifier: predicts hidden labels given observed sequence.
- Assigns a label to each region independently.
- Often used for OOV and error detection (Burget et al 08, Rastrow et al. 09, White et al 07, Hazen and Bassi 2001)
- Problems:
  - OOVs tend to be recognized as two or more words - OOV regions tend to co-occur (47% - 1 word, 40% - 2 words, 9% - 3 words, 4% - 4 or more).

Maximum Entropy Classifier



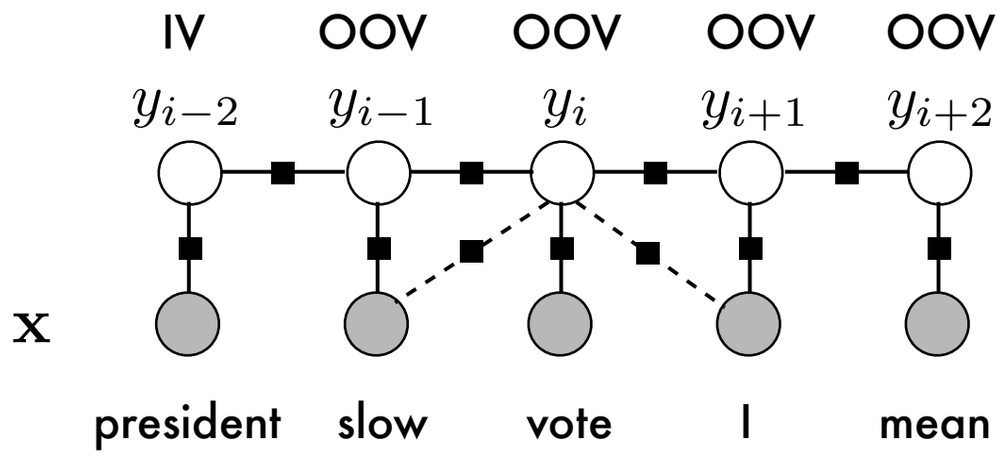
- Discriminative classifier: predicts hidden labels given observed sequence.
- Assigns a label to each region independently.
- Often used for OOV and error detection (Burget et al 08, Rastrow et al. 09, White et al 07, Hazen and Bassi 2001)
- Problems:
  - OOVs tend to be recognized as two or more words - OOV regions tend to co-occur (47% - 1 word, 40% - 2 words, 9% - 3 words, 4% - 4 or more).
  - Context can be helpful identifying OOVs. Perhaps they have specific distributional similarities (tend to be name-entities, rare nouns).

Conditional Random Fields



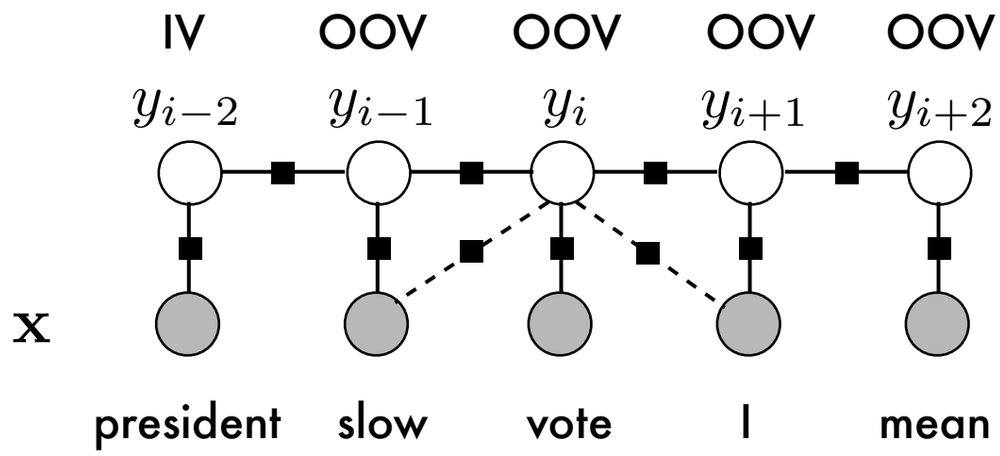
- Generalize MaxEnt models to sequence tasks.

Conditional Random Fields



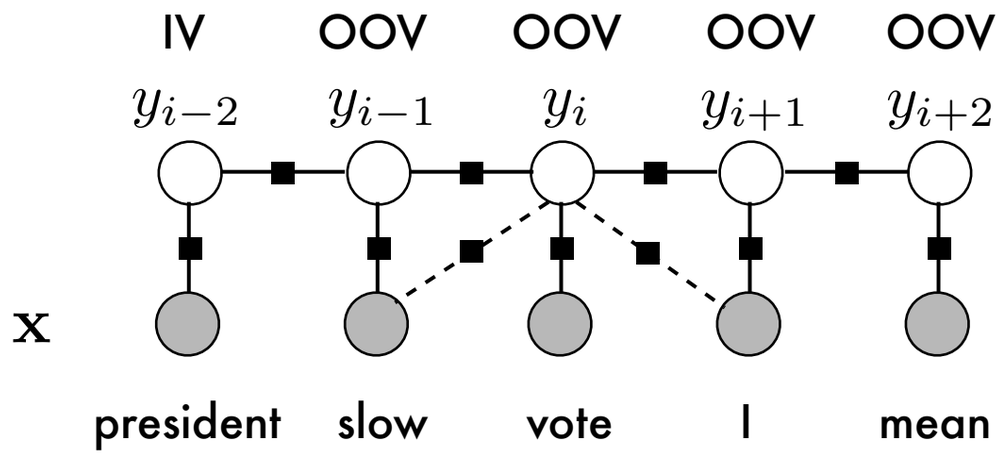
- Generalize MaxEnt models to sequence tasks.

Conditional Random Fields



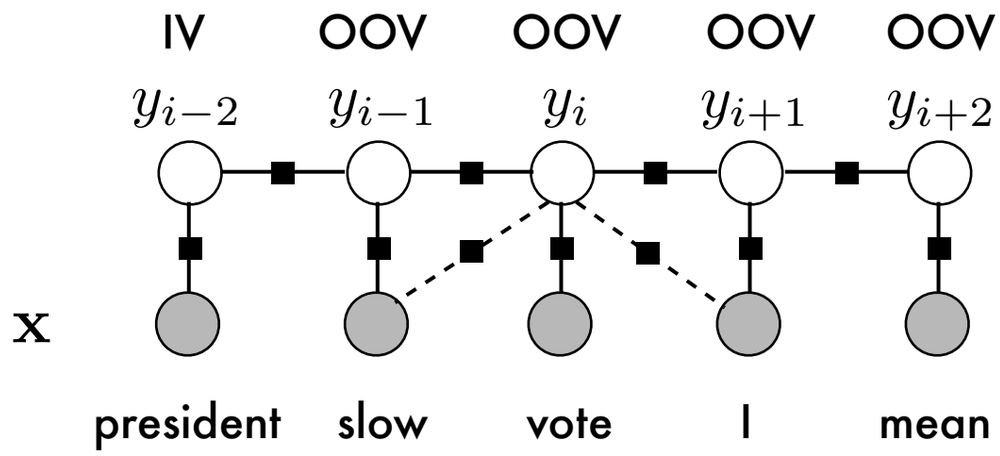
- Generalize MaxEnt models to sequence tasks.
- Sequence model: finds optimal labels for entire sequence instead of greedy local decisions.

Conditional Random Fields



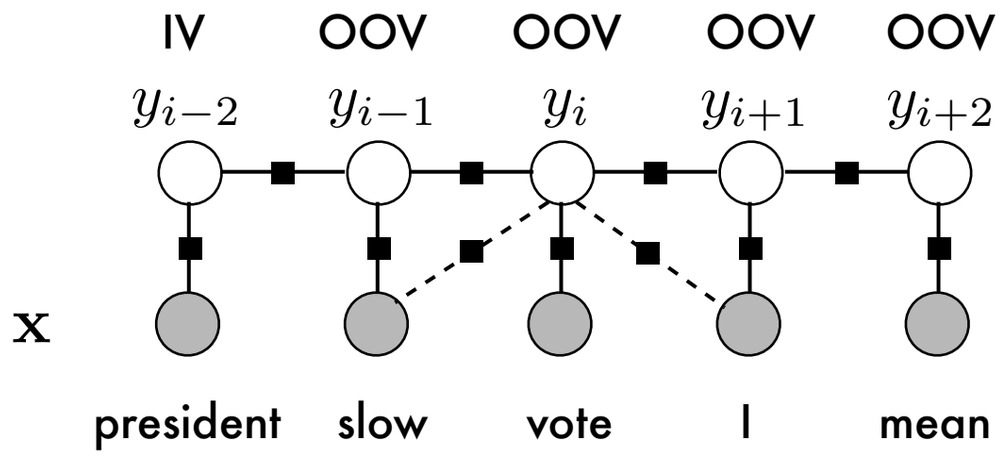
- Generalize MaxEnt models to sequence tasks.
- Sequence model: finds optimal labels for entire sequence instead of greedy local decisions.
- Used in Name-Entity-Recognition, POS Tagging (Pereira et al 1993), Sentence Boundary Detection (Liu et al. 2005).

Conditional Random Fields



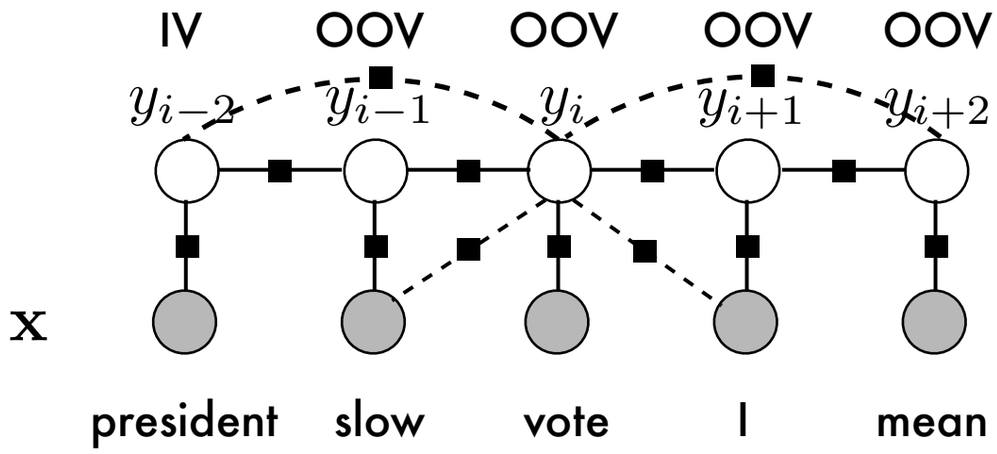
- Generalize MaxEnt models to sequence tasks.
- Sequence model: finds optimal labels for entire sequence instead of greedy local decisions.
- Used in Name-Entity-Recognition, POS Tagging (Pereira et al 1993), Sentence Boundary Detection (Liu et al. 2005).
- Linear-chain CRF: introduces dependencies between a label and its neighbors.

Conditional Random Fields



- Generalize MaxEnt models to sequence tasks.
- Sequence model: finds optimal labels for entire sequence instead of greedy local decisions.
- Used in Name-Entity-Recognition, POS Tagging (Pereira et al 1993), Sentence Boundary Detection (Liu et al. 2005).
- Linear-chain CRF: introduces dependencies between a label and its neighbors.
- More context in labels: Second-order CRF, BIO encoding.

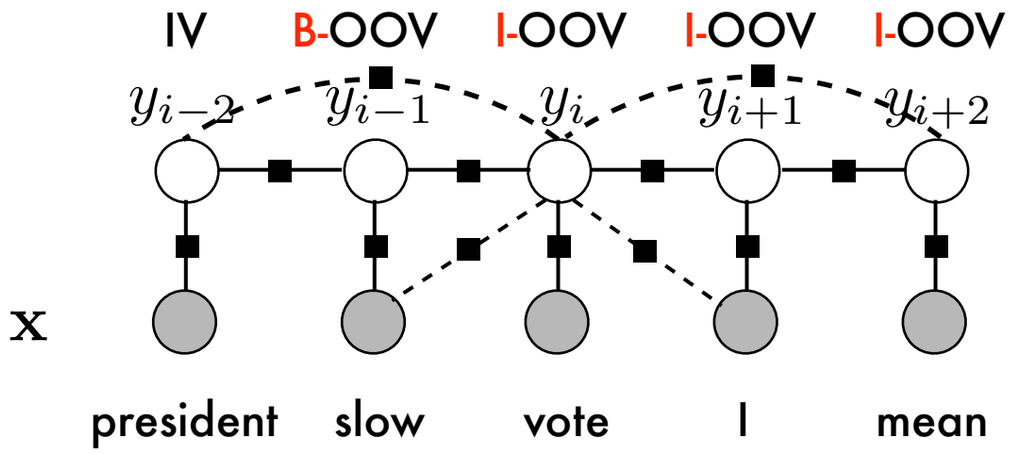
Conditional Random Fields



2nd order CRF

- Generalize MaxEnt models to sequence tasks.
- Sequence model: finds optimal labels for entire sequence instead of greedy local decisions.
- Used in Name-Entity-Recognition, POS Tagging (Pereira et al 1993), Sentence Boundary Detection (Liu et al. 2005).
- Linear-chain CRF: introduces dependencies between a label and its neighbors.
- More context in labels: Second-order CRF, BIO encoding.

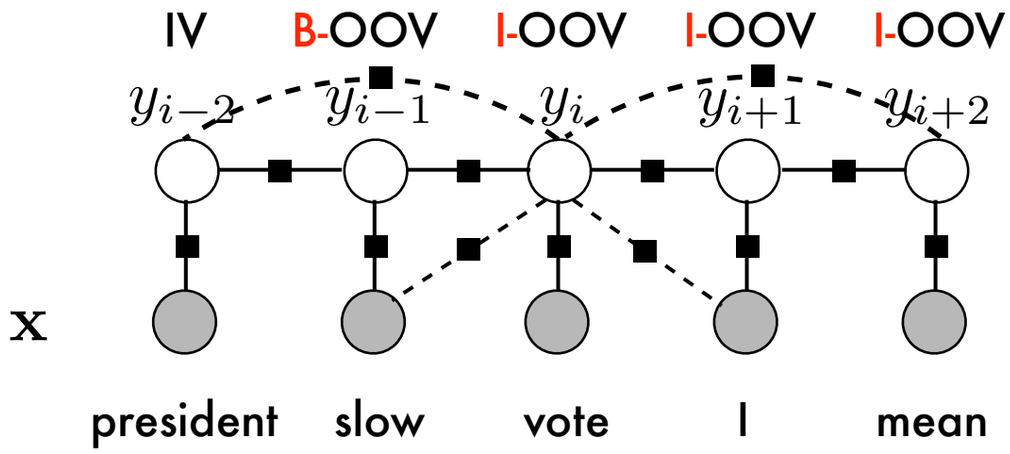
Conditional Random Fields



2nd order CRF
BIO encoding

- Generalize MaxEnt models to sequence tasks.
- Sequence model: finds optimal labels for entire sequence instead of greedy local decisions.
- Used in Name-Entity-Recognition, POS Tagging (Pereira et al 1993), Sentence Boundary Detection (Liu et al. 2005).
- Linear-chain CRF: introduces dependencies between a label and its neighbors.
- More context in labels: Second-order CRF, BIO encoding.

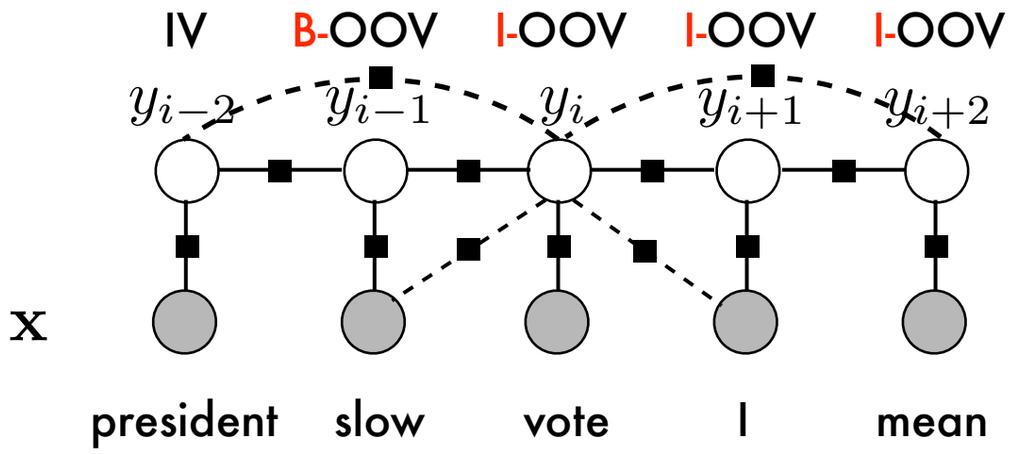
Conditional Random Fields



2nd order CRF
BIO encoding

- Generalize MaxEnt models to sequence tasks.
- Sequence model: finds optimal labels for entire sequence instead of greedy local decisions.
- Used in Name-Entity-Recognition, POS Tagging (Pereira et al 1993), Sentence Boundary Detection (Liu et al. 2005).
- Linear-chain CRF: introduces dependencies between a label and its neighbors.
- More context in labels: Second-order CRF, BIO encoding.

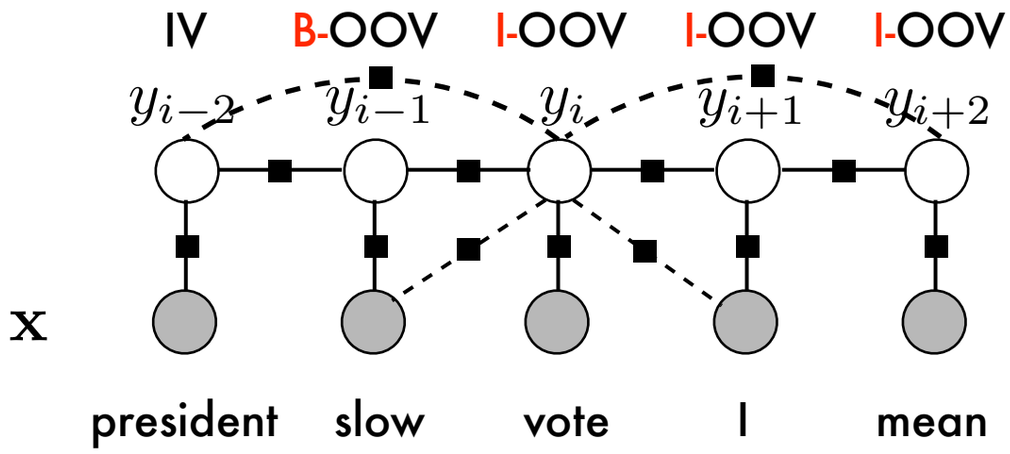
Conditional Random Fields



2nd order CRF
BIO encoding

- Can we improve OOV detection by framing it as a sequence labeling task?
- Does local and global information from observed sequence help?

Conditional Random Fields



2nd order CRF
BIO encoding

- Can we improve OOV detection by framing it as a sequence labeling task?
- Does local and global information from observed sequence help?

YES!

Outline

- Motivation
- Previous work in OOV detection
- Hybrid system for OOV detection (baseline)
- Overview: Conditional Random Fields vs Maximum Entropy
- Experimental Setup and Evaluation
- Improvements over baseline system
  - Generalize to sequence labeling problem
  - Incorporate local lexical information
  - Incorporate global information
- Conclusions and future work

Experimental Setup

- OOV Detection data-set: designed to emphasize OOVs (Can et. al. 2009)
  - 100 Hours (English Broadcast News)
  - 1290 unique OOVs, at least 5 instances per word, short OOVs excluded.
  - Divide 100 hrs into 5 hrs development set (used to train OOV-detector) and 95 hrs for testing.
  - 2% OOV rate on both development and test.

Experimental Setup

- OOV Detection data-set: designed to emphasize OOVs (Can et. al. 2009)
  - 100 Hours (English Broadcast News)
  - 1290 unique OOVs, at least 5 instances per word, short OOVs excluded.
  - Divide 100 hrs into 5 hrs development set (used to train OOV-detector) and 95 hrs for testing.
  - 2% OOV rate on both development and test.

- Training set LVCSR system:
  - IBM Speech Recognition Toolkit (Saltau et al. 2005)
  - Train on 300 Hours of English Broadcast News.
  - Language model 400M words, 83K word vocabulary.

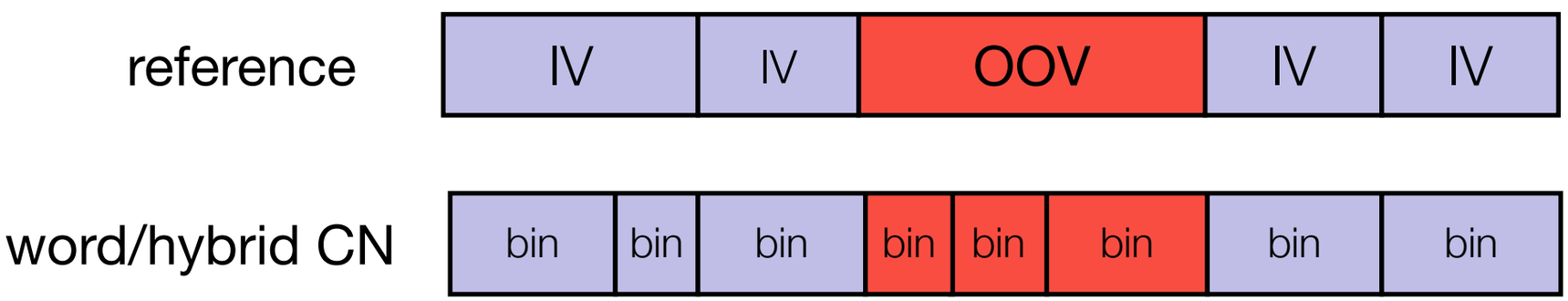
Experimental Setup

- OOV Detection data-set: designed to emphasize OOVs (Can et. al. 2009)
  - 100 Hours (English Broadcast News)
  - 1290 unique OOVs, at least 5 instances per word, short OOVs excluded.
  - Divide 100 hrs into 5 hrs development set (used to train OOV-detector) and 95 hrs for testing.
  - 2% OOV rate on both development and test.

- Training set LVCSR system:
  - IBM Speech Recognition Toolkit (Saltau et al. 2005)
  - Train on 300 Hours of English Broadcast News.
  - Language model 400M words, 83K word vocabulary.

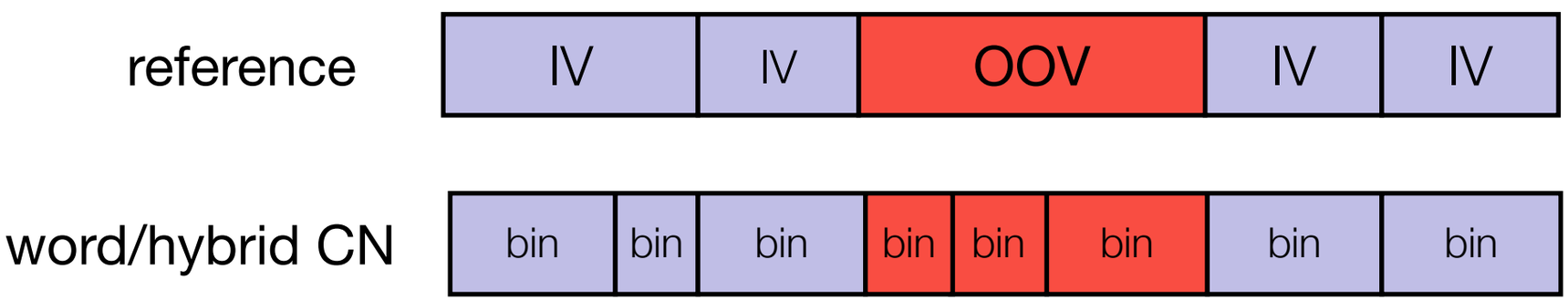
- Hybrid System:
  - 83K words and 20K fragments.
  - 1290 unique words are OOVs for both word and hybrid system.

Evaluation



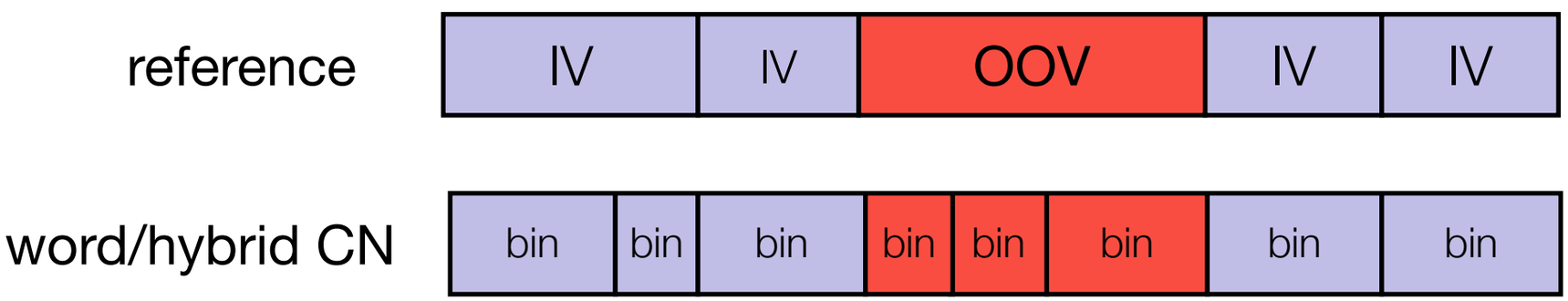
- The ASR transcript is aligned to the reference at the confusion bin level.

Evaluation



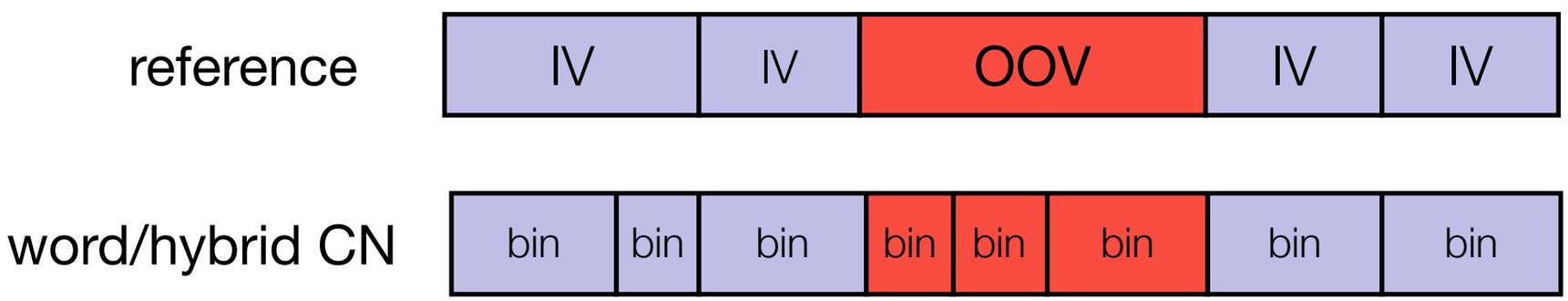
- The ASR transcript is aligned to the reference at the confusion bin level.
- Each bin is assigned a score obtained form the MaxEnt or CRF.

Evaluation



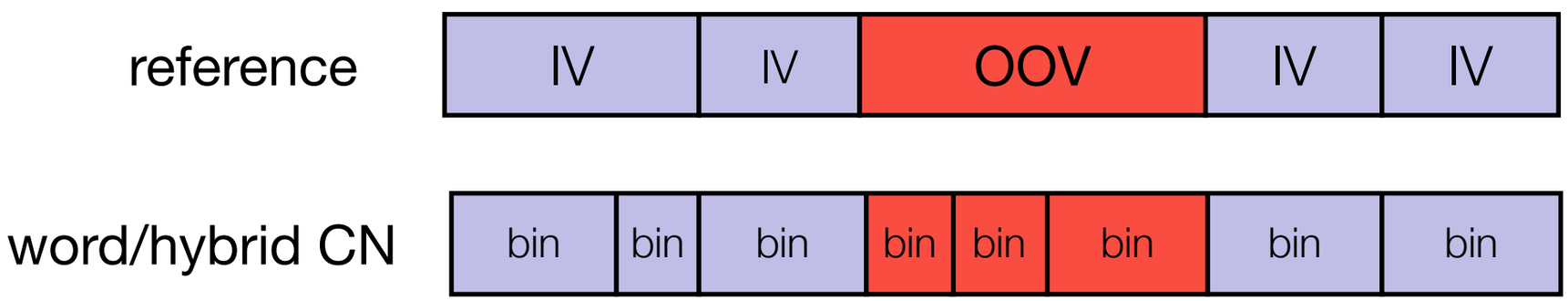
- The ASR transcript is aligned to the reference at the confusion bin level.
- Each bin is assigned a score obtained form the MaxEnt or CRF.
- A threshold for the scored is varied to generate IV and OOV tags.

Evaluation



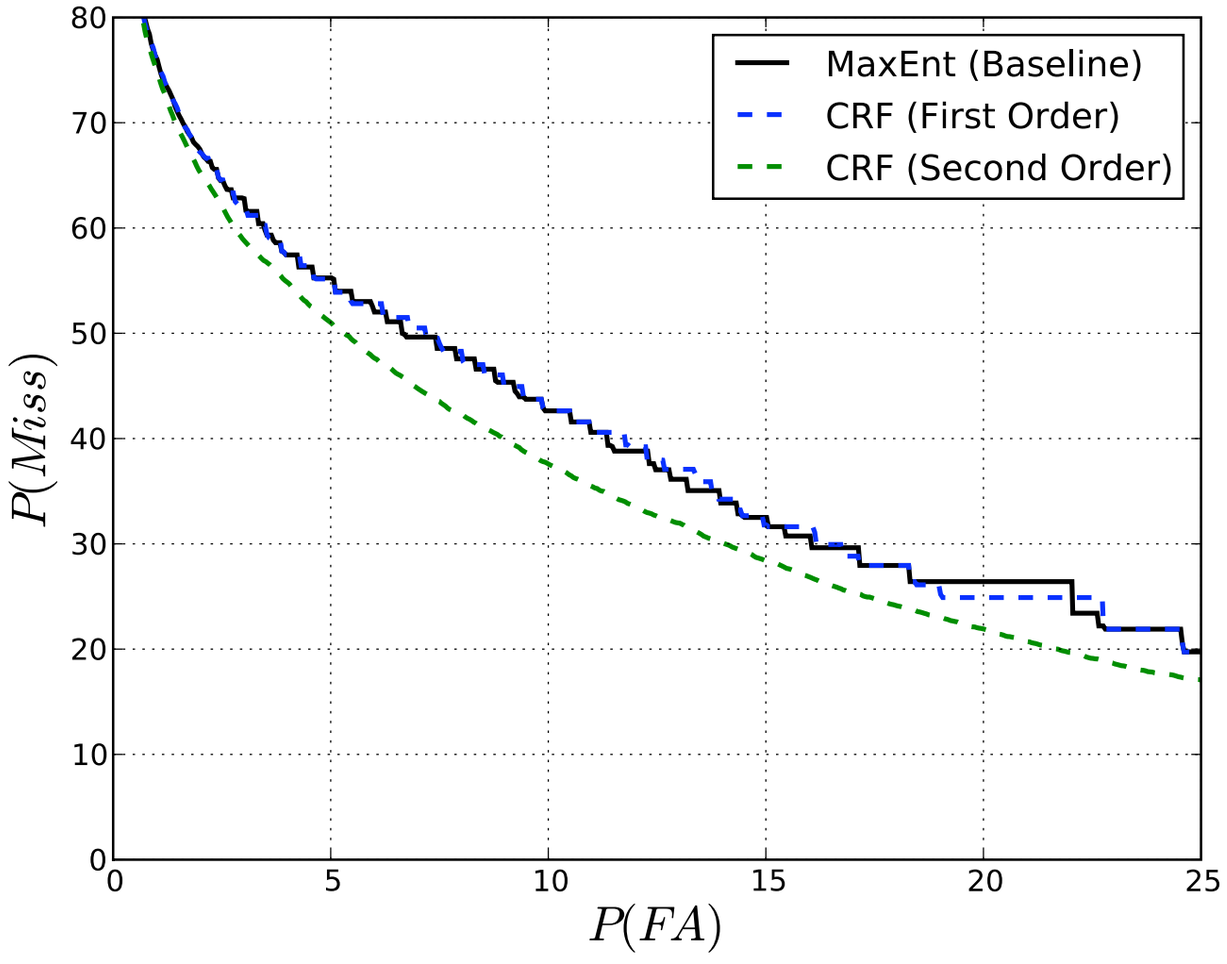
- The ASR transcript is aligned to the reference at the confusion bin level.
- Each bin is assigned a score obtained form the MaxEnt or CRF.
- A threshold for the scored is varied to generate IV and OOV tags.
- Present results in terms of false-alarms and miss probabilities using detection error tradeoff (DET) curve.

Evaluation



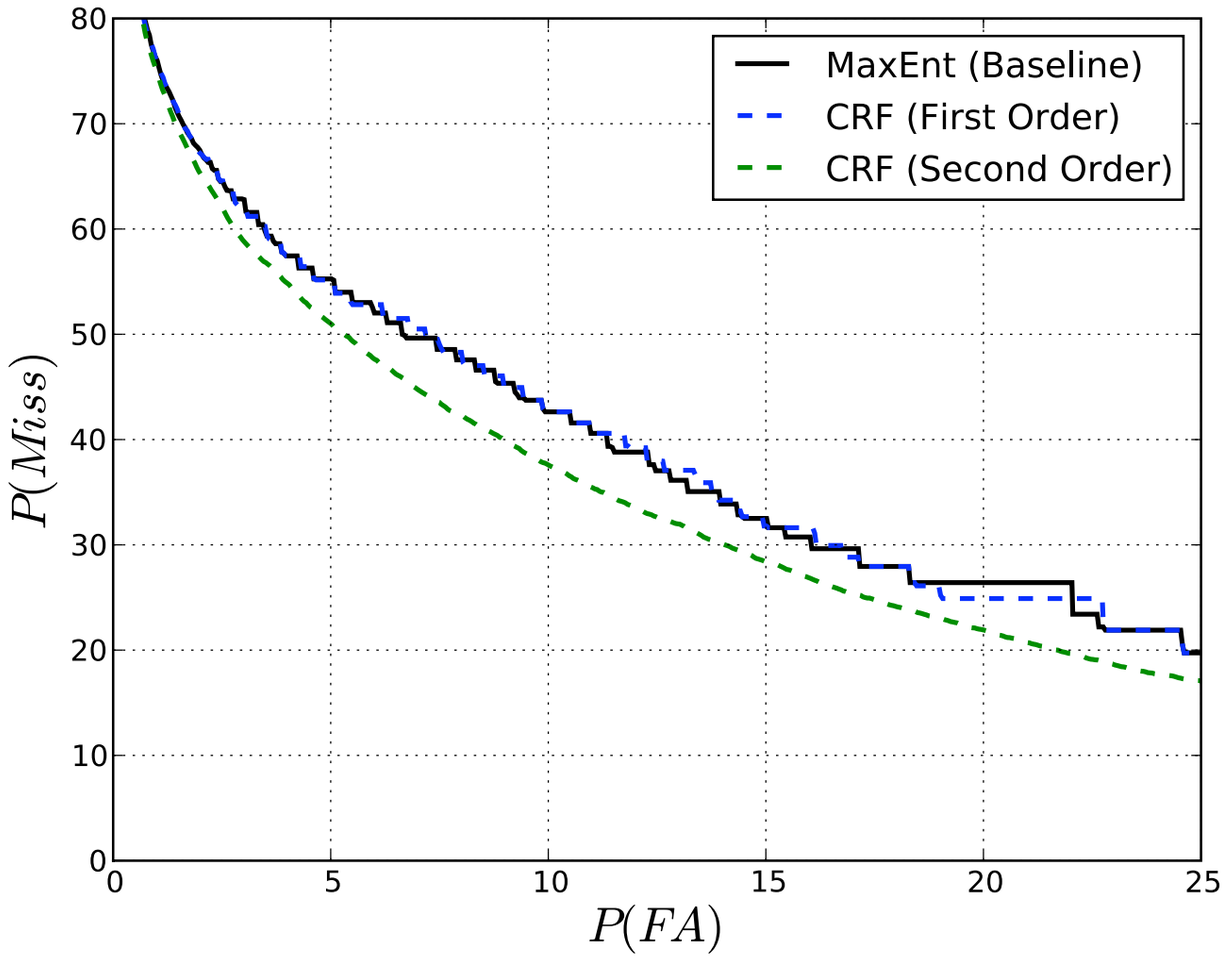
- The ASR transcript is aligned to the reference at the confusion bin level.
- Each bin is assigned a score obtained form the MaxEnt or CRF.
- A threshold for the scored is varied to generate IV and OOV tags.
- Present results in terms of false-alarms and miss probabilities using detection error tradeoff (DET) curve.
- Present results on *un-observed* OOVs (not on vocabulary of ASR or training set for OOV-detector).

Result: MaxEnt vs CRF model



- Baseline features:
  - Word-Entropy
  - Fragment-Posterior
- Features quantized using uniform partitioning (50 bins, minimum 100 samples per bin)
- 5% absolute improvement with same features, at 10% FA rate.

Result: MaxEnt vs CRF model



- Baseline features:
  - Word-Entropy
  - Fragment-Posterior
- Features quantized using uniform partitioning (50 bins, minimum 100 samples per bin)
- 5% absolute improvement with same features, at 10% FA rate.

Didn't help:

- Higher Order CRF

Local Lexical Context

- Common in sequence models: include features from local lexical context: e.g. in Name Entity Recognition: Mr. Milosevic.

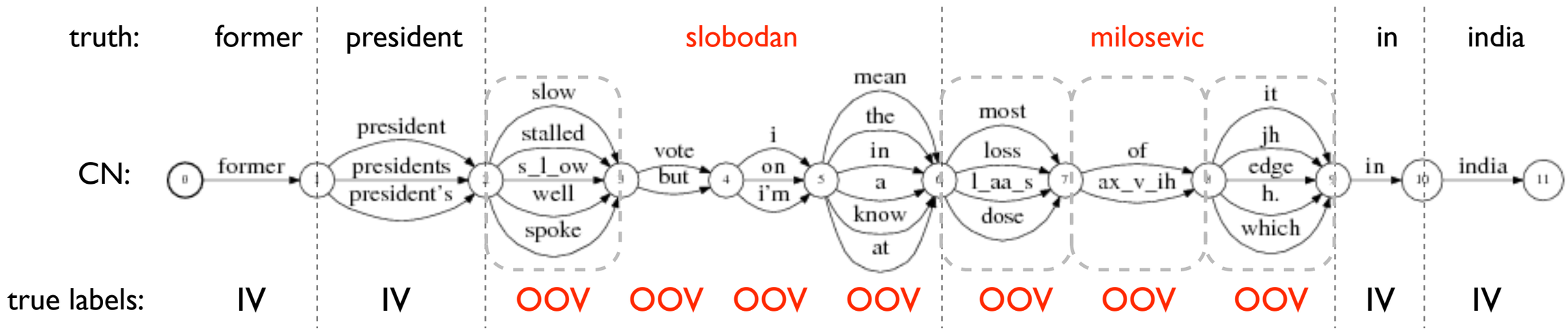
Local Lexical Context

- Common in sequence models: include features from local lexical context: e.g. in Name Entity Recognition: Mr. Milosevic.
- Use best-hypothesis from ASR: “former president“ likely to be followed by “of” or <name>.
- Features of the form:

$$\text{current\_word} = X$$

$$\text{word}[-2]/\text{word}[-1] = \text{former}/\text{president}$$

Local Lexical Context

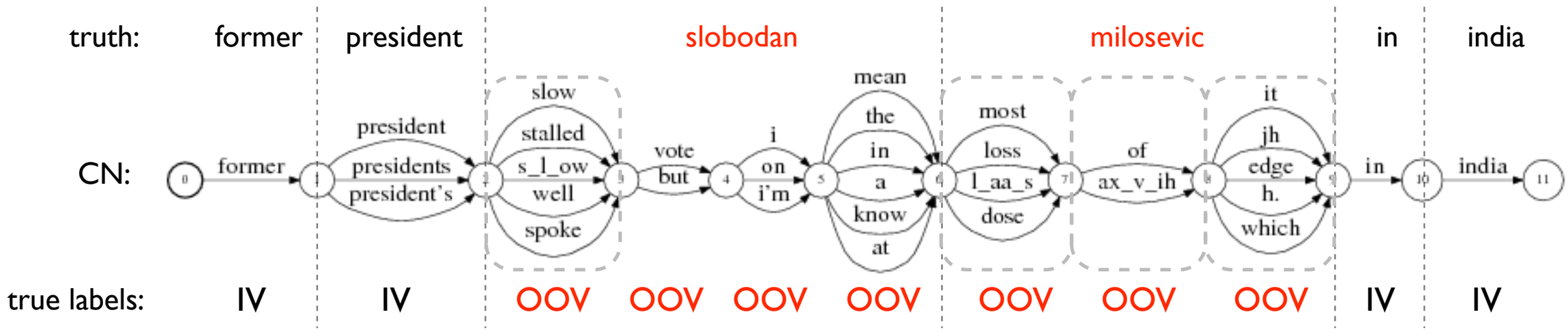


- Common in sequence models: include features from local lexical context: e.g. in Name Entity Recognition: Mr. Milosevic.
- Use best-hypothesis from ASR: "former president" likely to be followed by "of" or <name>.
- Features of the form:

$$\text{current\_word} = X$$

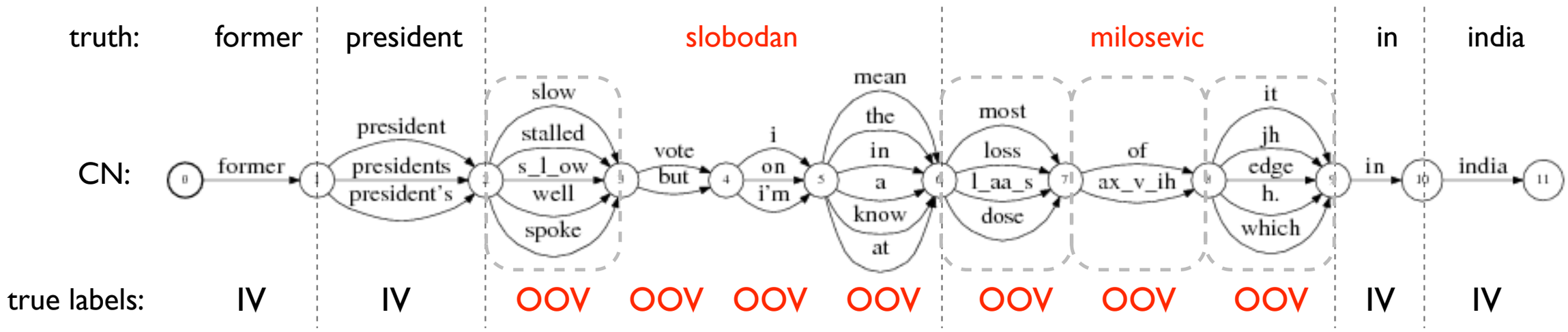
$$\text{word}[-2]/\text{word}[-1] = \text{former}/\text{president}$$

Local Lexical Context



- Common in sequence models: include features from local lexical context: e.g. in Name Entity Recognition: Mr. Milosevic.
- Use best-hypothesis from ASR: “former president“ likely to be followed by “of” or <name>.

Local Lexical Context

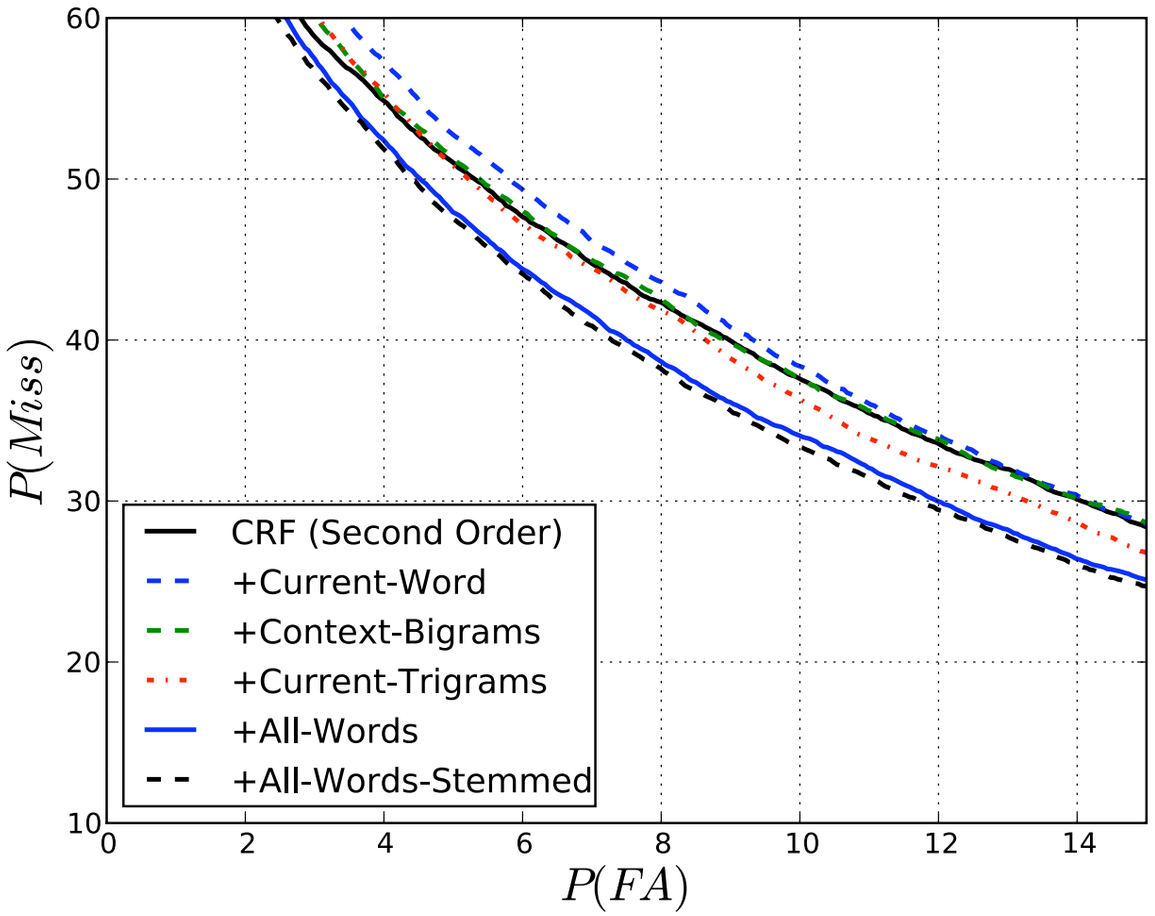


- Common in sequence models: include features from local lexical context: e.g. in Name Entity Recognition: Mr. Milosevic.
- Use best-hypothesis from ASR: "former president" likely to be followed by "of" or <name>.
- Features of the form:

$$\text{current\_word} = X$$

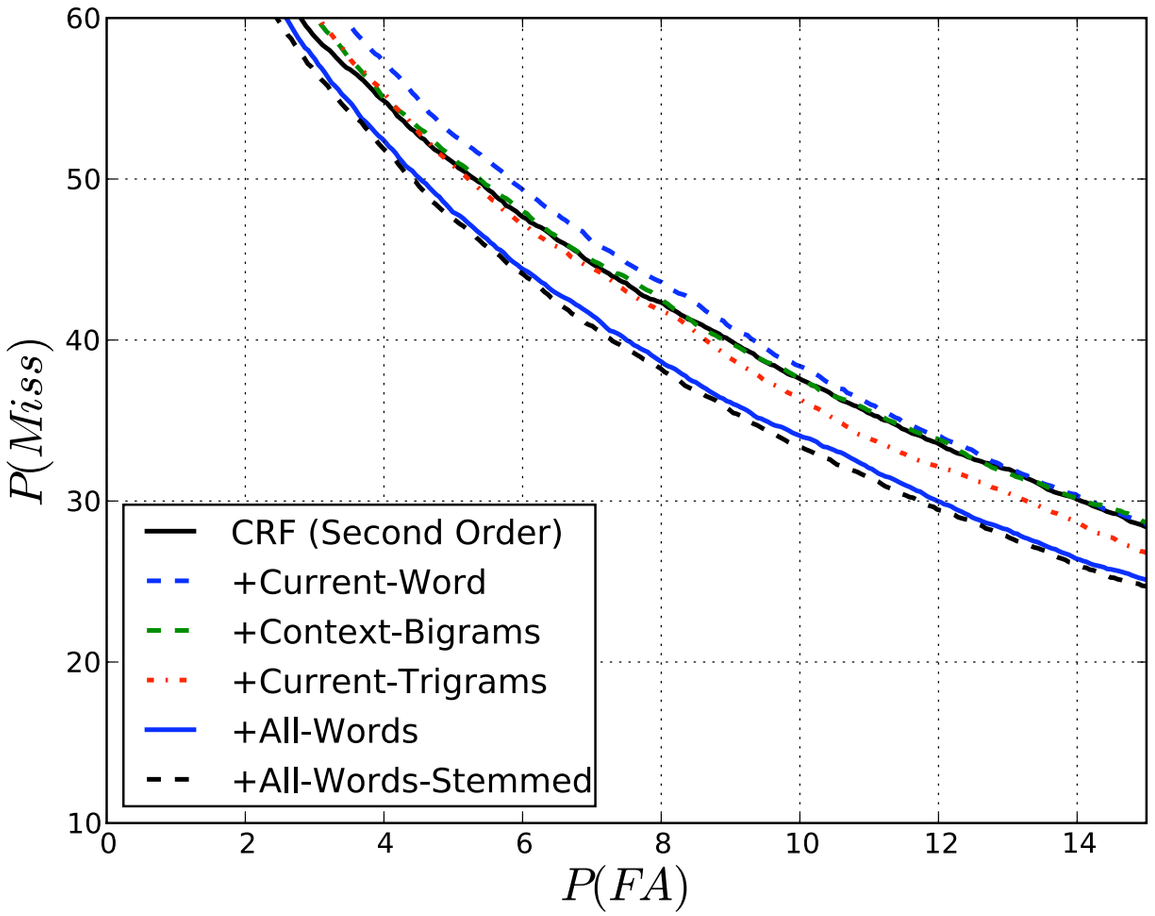
$$\text{word}[-2]/\text{word}[-1] = \text{former}/\text{president}$$

Result: local lexical context



- Features used:
  - Current-Word
  - Context Bigrams: bigrams from 5-word window (ignore current-bin)
  - Current-Trigrams: trigrams including current bin in 5-word window
  - All Words: All above features
  - All Words Stemmed.

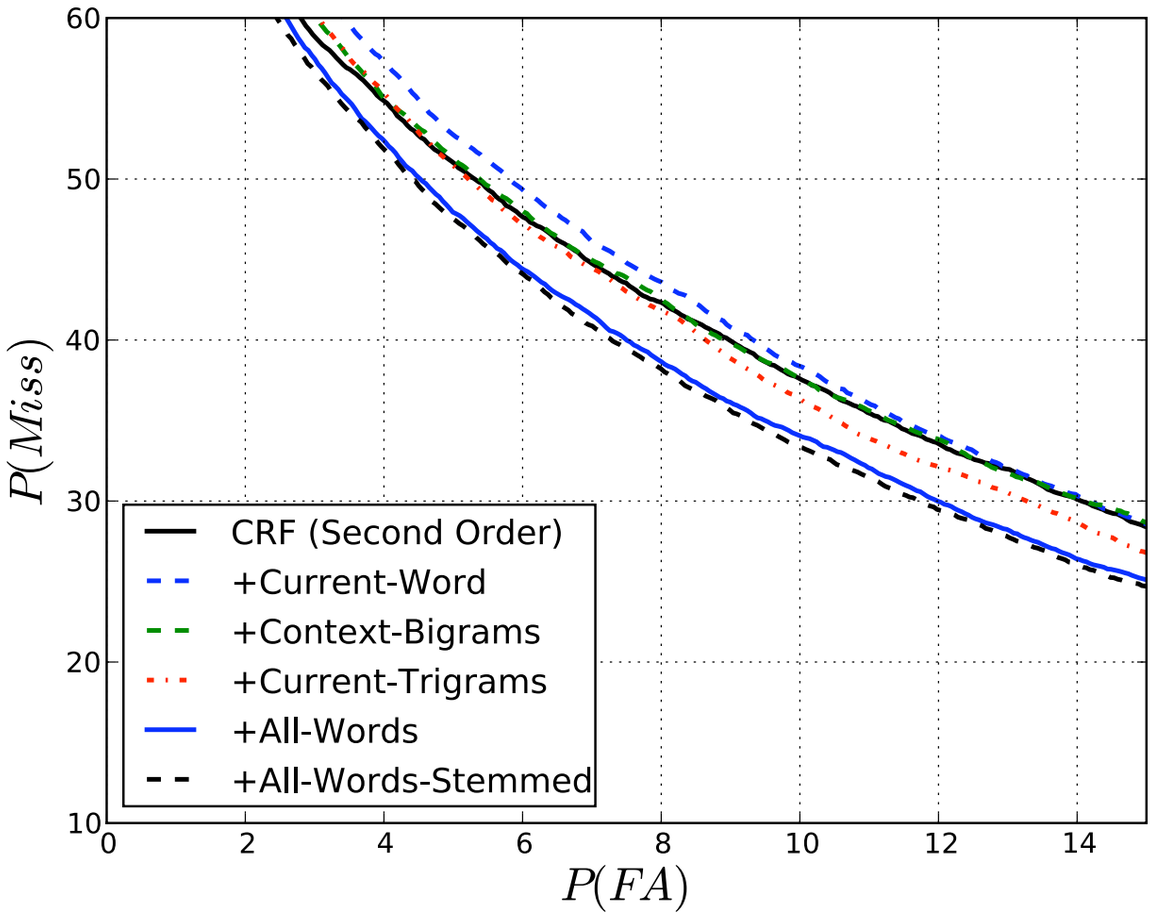
Result: local lexical context



- Features used:
  - Current-Word
  - Context Bigrams: bigrams from 5-word window (ignore current-bin)
  - Current-Trigrams: trigrams including current bin in 5-word window
  - All Words: All above features
  - All Words Stemmed.

- 4.2% absolute improvement over previous result (9.3% over baseline) at 10% FA rate.

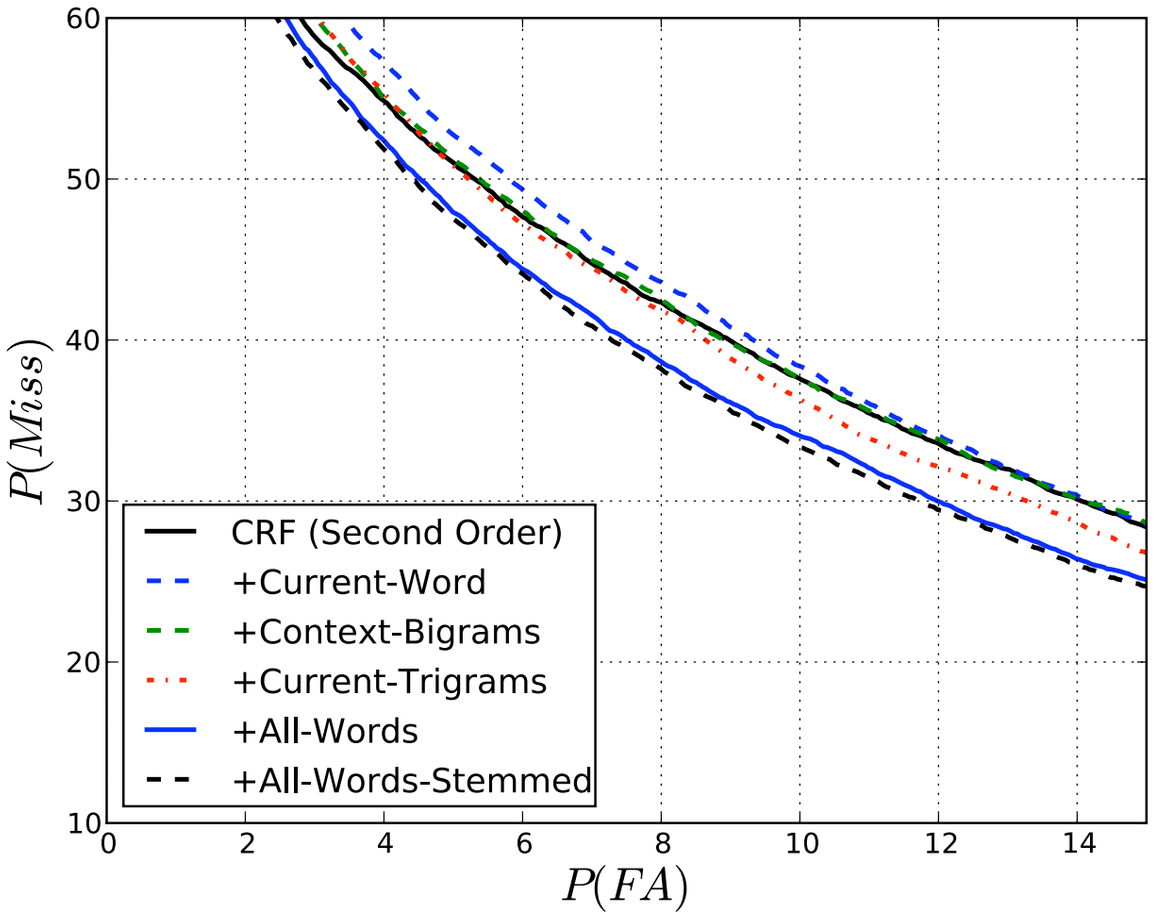
Result: local lexical context



- Features used:
  - Current-Word
  - Context Bigrams: bigrams from 5-word window (ignore current-bin)
  - Current-Trigrams: trigrams including current bin in 5-word window
  - All Words: All above features
  - All Words Stemmed.

- 4.2% absolute improvement over previous result (9.3% over baseline) at 10% FA rate.
- Combining context and current bin achieves most of the gain.

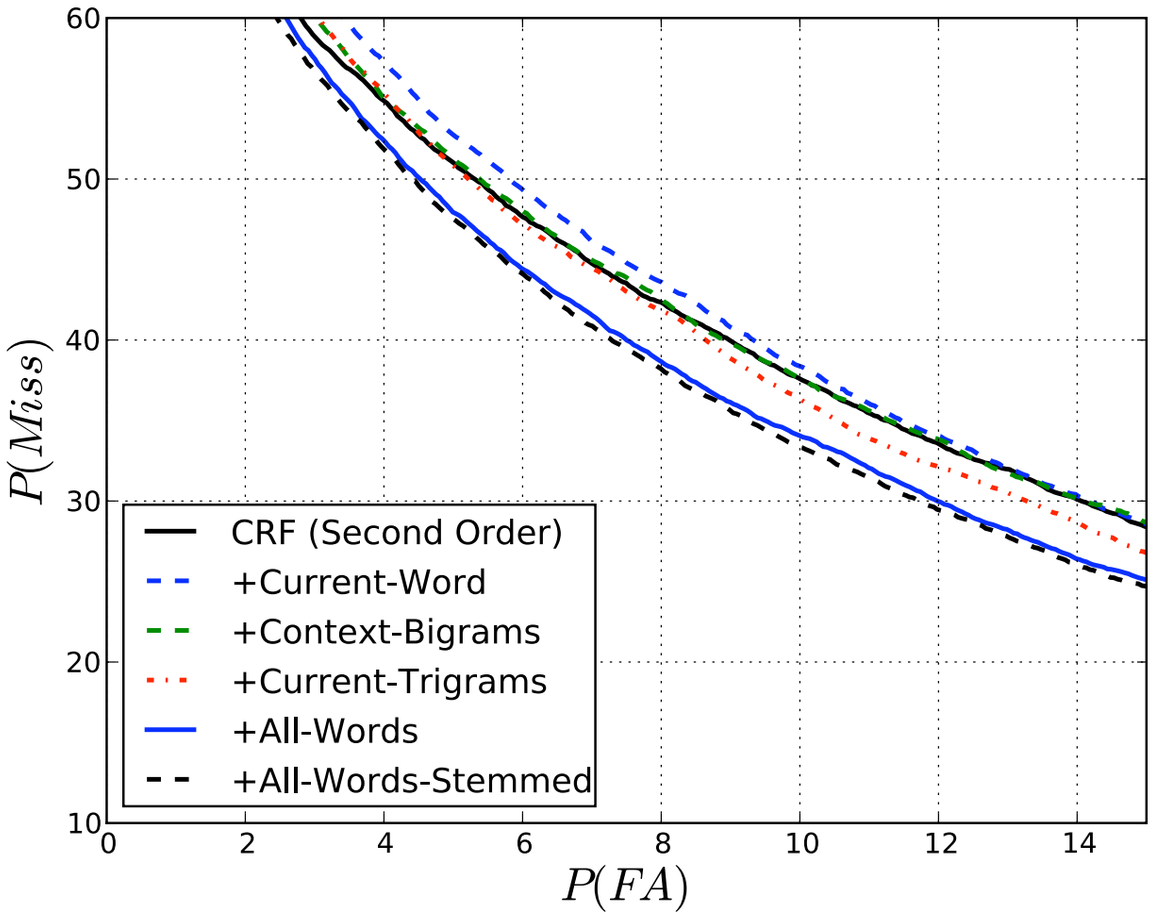
Result: local lexical context



- Features used:
  - Current-Word
  - Context Bigrams: bigrams from 5-word window (ignore current-bin)
  - Current-Trigrams: trigrams including current bin in 5-word window
  - All Words: All above features
  - All Words Stemmed.

- 4.2% absolute improvement over previous result (9.3% over baseline) at 10% FA rate.
- Combining context and current bin achieves most of the gain.
- Indicates OOVs tend to occur with certain distributional characteristics.

Result: local lexical context



Didn't help:

- Adding substrings from current and context.
- Baseline features from neighboring bins.

- Features used:
  - Current-Word
  - Context Bigrams: bigrams from 5-word window (ignore current-bin)
  - Current-Trigrams: trigrams including current bin in 5-word window
  - All Words: All above features
  - All Words Stemmed.

- 4.2% absolute improvement over previous result (9.3% over baseline) at 10% FA rate.
- Combining context and current bin achieves most of the gain.
- Indicates OOVs tend to occur with certain distributional characteristics.

Incorporating Global Context

- We include information from entire utterance using LM and POS tagging.

Incorporating Global Context

- We include information from entire utterance using LM and POS tagging.
  - The probability of an utterance as given by an Language Model is a measure of its fluency.

Incorporating Global Context

- We include information from entire utterance using LM and POS tagging.
  - The probability of an utterance as given by an Language Model is a measure of its fluency.
  - OOVs tend to take specific syntactic roles (over 50% are proper nouns).

Incorporating Global Context

- We include information from entire utterance using LM and POS tagging.
  - The probability of an utterance as given by an Language Model is a measure of its fluency.
  - OOVs tend to take specific syntactic roles (over 50% are proper nouns).
- Use features derived from N-gram and syntactic Language Models.

Incorporating Global Context

- We include information from entire utterance using LM and POS tagging.
  - The probability of an utterance as given by an Language Model is a measure of its fluency.
  - OOVs tend to take specific syntactic roles (over 50% are proper nouns).
- Use features derived from N-gram and syntactic Language Models.

N-gram Language Model

$$
\begin{aligned}
P(w_1^m) &= \prod_{i=1}^{m} P(w_i | w_1^{i-1}) \\
&\approx \prod_{i=1}^{m} P(w_i | w_{i-N+1} \dots w_{i-1})
\end{aligned}
$$

Syntactic Language Model

- A Joint Language Model with Fine-Grained Syntactic tags (Filimonov and Harper 2009)
- Estimates joint probability of word and its syntactic tag .

$$P(w_1^m t_1^m) = \prod_{i=1}^{m} P(w_i t_i | w_1^{i-1} t_1^{i-1})$$

Syntactic Language Model

- A Joint Language Model with Fine-Grained Syntactic tags (Filimonov and Harper 2009)
- Estimates joint probability of word and its syntactic tag .

$$P(w_1^m t_1^m) = \prod_{i=1}^{m} P(w_i t_i | w_1^{i-1} t_1^{i-1})$$

Syntactic Language Model

- A Joint Language Model with Fine-Grained Syntactic tags (Filimonov and Harper 2009)
- Estimates joint probability of word and its syntactic tag .

$$P(w_1^m t_1^m) = \prod_{i=1}^{m} P(w_i t_i | w_1^{i-1} t_1^{i-1})$$

$$P(w_1^m) = \sum_{t_1 \ldots t_m} \prod_{i=1}^{m} P(w_i t_i | w_1^{i-1} t_1^{i-1})$$

$$\approx \sum_{t_1 \ldots t_m} \prod_{i=1}^{m} P(w_i t_i | w_{i-n+1}^{i-1} t_{i-n+1}^{i-1})$$

Syntactic Language Model

- A Joint Language Model with Fine-Grained Syntactic tags (Filimonov and Harper 2009)
- Estimates joint probability of word and its syntactic tag .

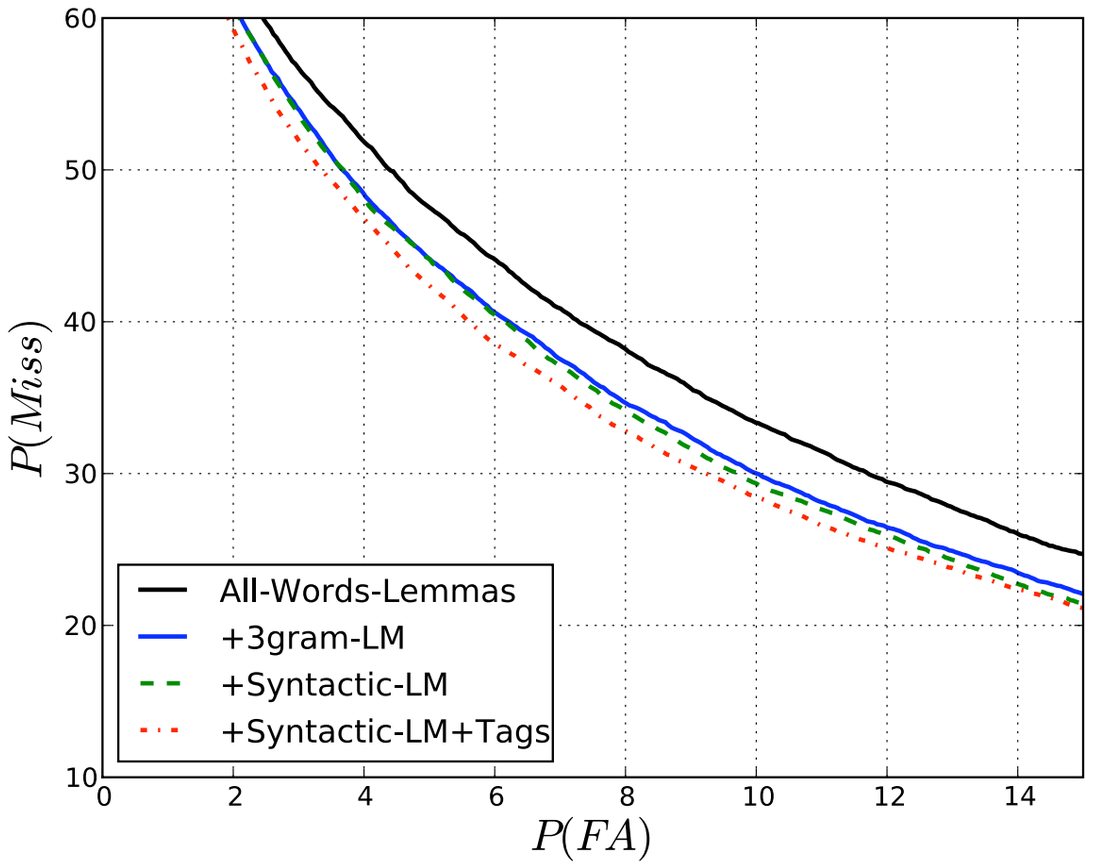
$$P(w_1^m t_1^m) = \prod_{i=1}^{m} P(w_i t_i | w_1^{i-1} t_1^{i-1})$$

$$P(w_1^m) = \sum_{t_1 \ldots t_m} \prod_{i=1}^{m} P(w_i t_i | w_1^{i-1} t_1^{i-1})$$

$$\approx \sum_{t_1 \ldots t_m} \prod_{i=1}^{m} P(w_i t_i | w_{i-n+1}^{i-1} t_{i-n+1}^{i-1})$$

- Tags are extracted from parse trees and include: word POS-tag, label of immediate parent, and relative position of word among its siblings.
- Hidden states carry global information since it estimates most likely tag sequence for entire utterance.

Result: global context (LMs)



- Features used:
  - Likelihood ratio:

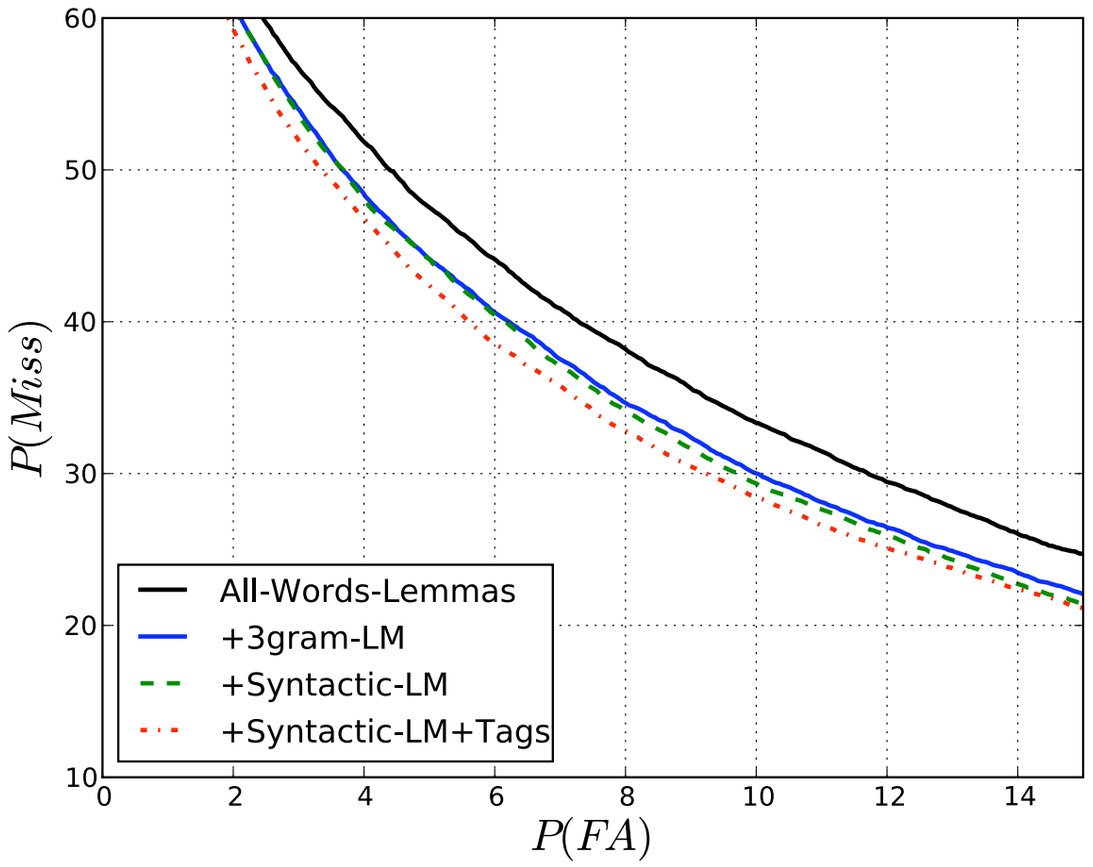
$$\log \frac{p(utt)}{p(utt|w_i = unknown)}$$

  - Probability of utterance:

$$\frac{\log p(utt)}{length(utt)}$$

  - POS Tags in 5-tag window

Result: global context (LMs)



- Features used:
  - Likelihood ratio:

$$\log \frac{p(utt)}{p(utt|w_i = unknown)}$$

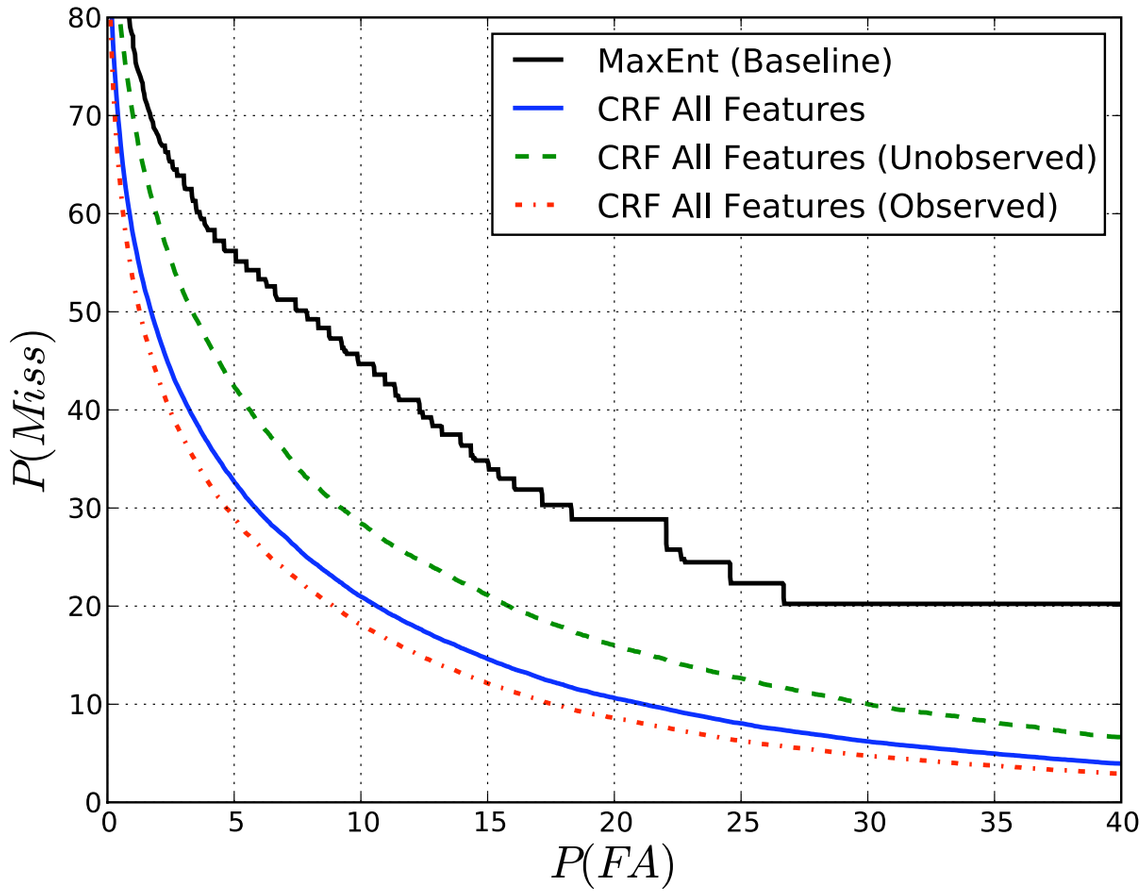
  - Probability of utterance:

$$\frac{\log p(utt)}{length(utt)}$$

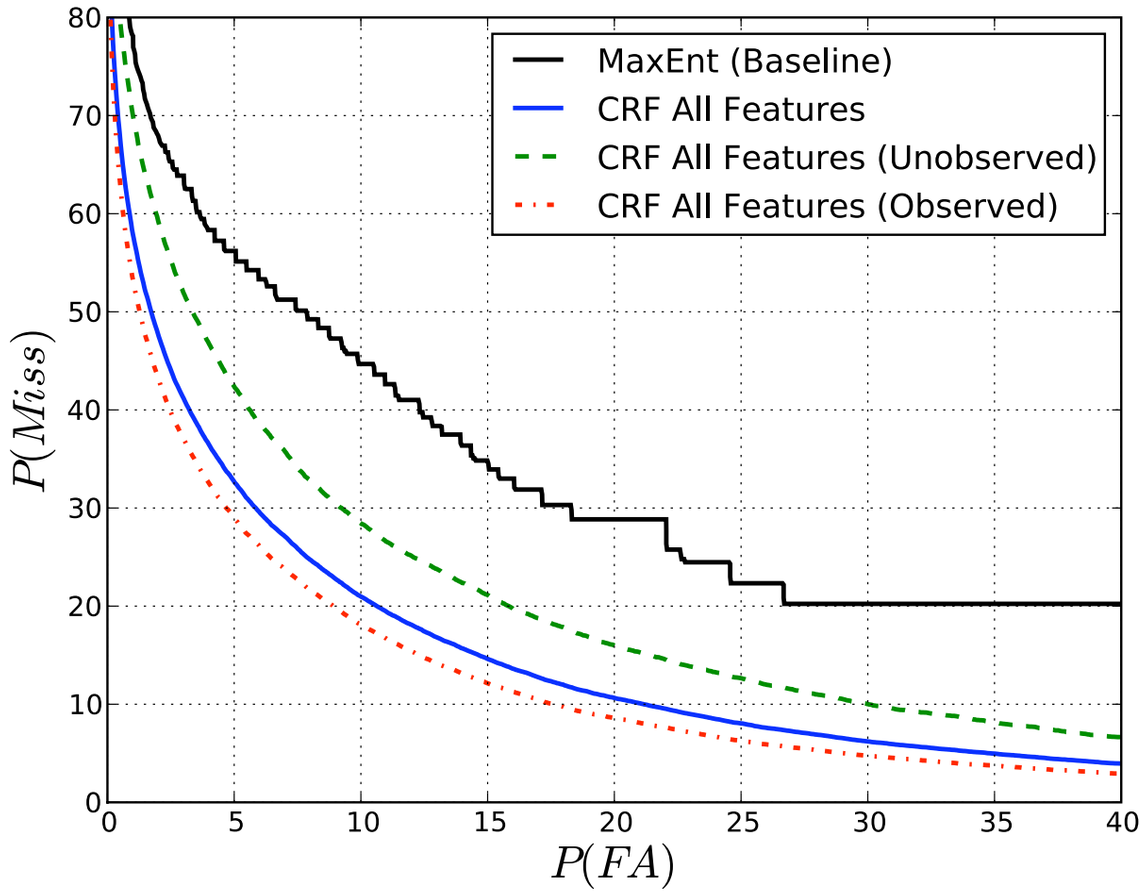
  - POS Tags in 5-tag window

- 4.9% absolute improvement over previous best result (14.2% over baseline) at 10% FA rate.
- Most of the additional gain achieved by N-gram Language model.

Final System

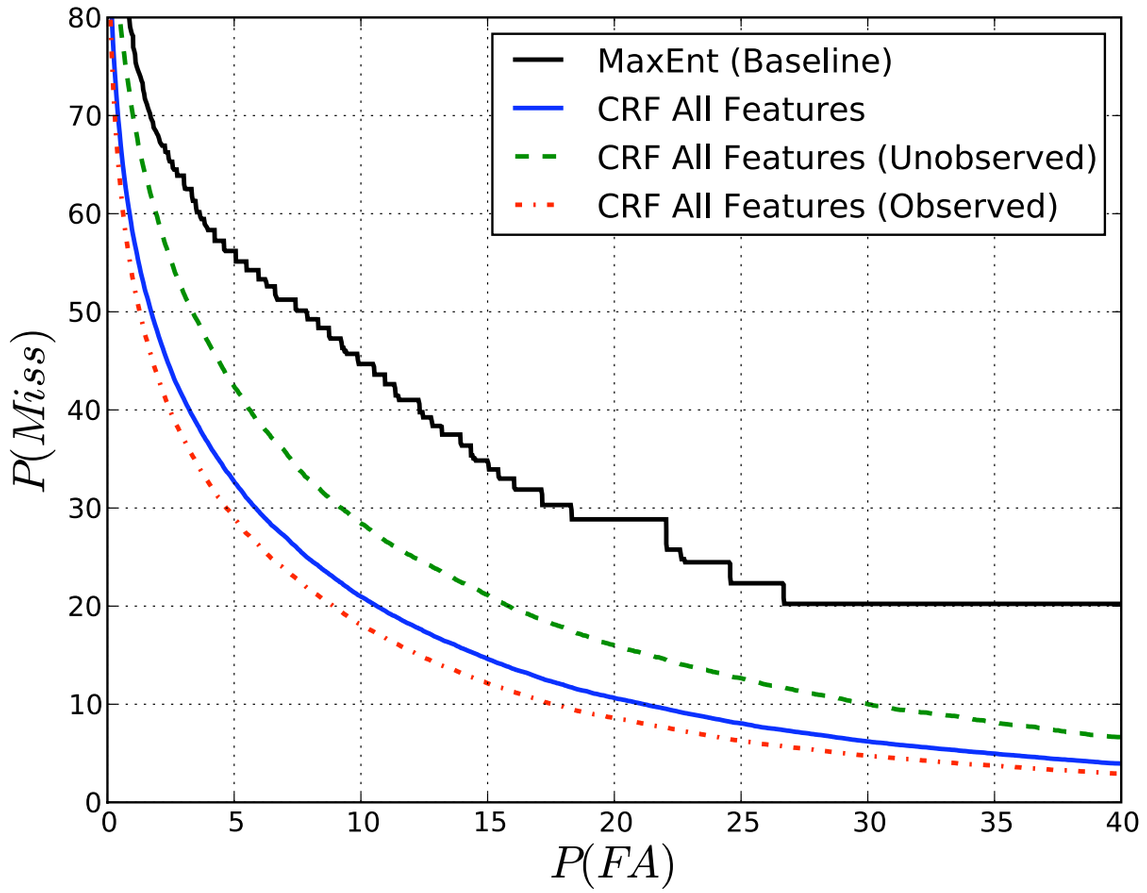


Final System



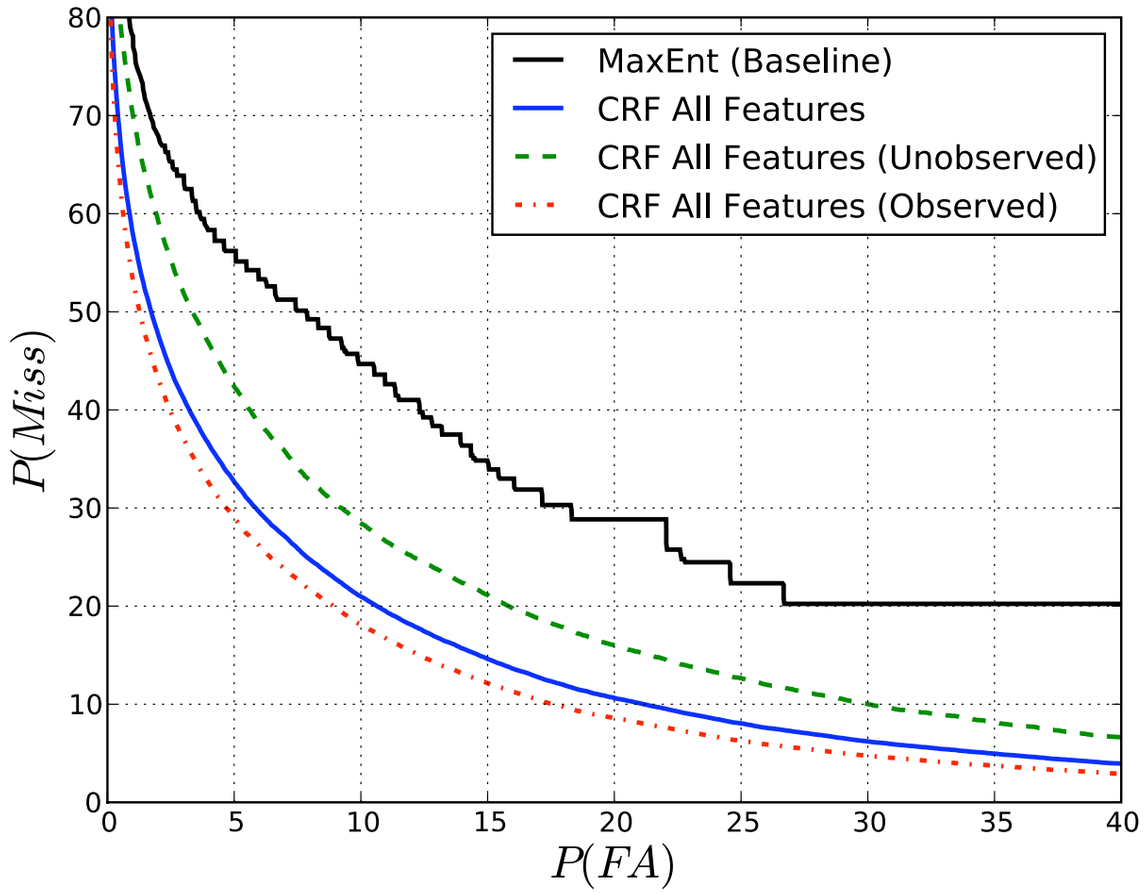
- Un-observed OOVs, at 10% FA rate, reduce miss OOV rate from 42.6% to 28.3%, 33.3% relative improvement.

Final System



- Un-observed OOVs, at 10% FA rate, reduce miss OOV rate from 42.6% to 28.3%, 33.3% relative improvement.
- Un-observed vs observed OOVs identical performance for baseline system. We achieve 55.6% relative when considering all OOVs.

Final System



- Un-observed OOVs, at 10% FA rate, reduce miss OOV rate from 42.6% to 28.3%, 33.3% relative improvement.
- Un-observed vs observed OOVs identical performance for baseline system. We achieve 55.6% relative when considering all OOVs.
- Baseline system flattens out at 26% FA rate, while CRF continues to decrease: useful if misses are more heavily penalized.

Conclusions

- Context helps detect OOVs! (submitted NAACL HLT 2010)
  - Integrating local lexical and global information helps.
  - Using sequence models (such as CRF) improves over a MaxEnt model, which treats problem as sequence of independent binary classification problems.

Conclusions

- Context helps detect OOVs! (submitted NAACL HLT 2010)
  - Integrating local lexical and global information helps.
  - Using sequence models (such as CRF) improves over a MaxEnt model, which treats problem as sequence of independent binary classification problems.
- At 10% FAs, reduced missed OOV rate from 42.6% to 28.4%, a 33% relative improvement.

Conclusions

- Context helps detect OOVs! (submitted NAACL HLT 2010)
  - Integrating local lexical and global information helps.
  - Using sequence models (such as CRF) improves over a MaxEnt model, which treats problem as sequence of independent binary classification problems.
- At 10% FAs, reduced missed OOV rate from 42.6% to 28.4%, a 33% relative improvement.
- Approach can be easily integrated with other proposed confidence measures, to enhance performance.

Future Work

- Include features from other hypothesis output by the recognizer (with associated confidence).
- Evaluate effect of OOV detector improvement in downstream applications such as Spoken Term Detection of OOVs.
- Context helps detect an OOV! Can we use it for detecting type of OOV? e.g. [person], [location], [organization], [other].

Thank you

- Collaborators: Mark Dredze, Denis Filimonov, Fred Jelinek.
- Ariya Rastrow and Chris White (supplying code for baseline).
- Abhinav Sethy and Bhuvana Ramabhadran (IBM) provided data and comments.

Questions?

Maximum Entropy vs CRF

MaxEnt

$$p(y_i|x) = \frac{\exp\left(\sum_{i=1}^{F} \lambda_i f_i(x, y_i)\right)}{\sum_{y'} \exp\left(\sum_{i=1}^{F} \lambda_i f_i(x, y')\right)}$$

CRF-linear

$$p(y_i|x) = \frac{1}{Z(x)} \exp\left\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x)\right\}$$

Automatic Speech Recognition (ASR)

- Find most likely word sequence uttered by a speaker given the acoustics.

Automatic Speech Recognition (ASR)

- Find most likely word sequence uttered by a speaker given the acoustics.

$$\hat{W} = \arg\max_{W} P(W|A)$$

Automatic Speech Recognition (ASR)

- Find most likely word sequence uttered by a speaker given the acoustics.

$$\hat{W} = \arg\max_W P(W|A) = \arg\max_W P(W)P(A|W)$$

Automatic Speech Recognition (ASR)

- Find most likely word sequence uttered by a speaker given the acoustics.

language model

acoustic model

$$\hat{W} = \arg\max_W P(W|A) = \arg\max_W P(W)P(A|W)$$

search

Automatic Speech Recognition (ASR)

- Find most likely word sequence uttered by a speaker given the acoustics.

language model

acoustic model

$$\hat{W} = \arg\max_W P(W|A) = \arg\max_W P(W)P(A|W)$$

search

- Assumptions: w belongs to a finite vocabulary (typically ~85K).

Automatic Speech Recognition (ASR)

- Find most likely word sequence uttered by a speaker given the acoustics.

language model

acoustic model

$$\hat{W} = \arg\max_W P(W|A) = \arg\max_W P(W)P(A|W)$$

search

- Assumptions: w belongs to a finite vocabulary (typically ~85K).
- Even with a large vocabulary, ASR systems will encounter words not seen during training: Out-Of-Vocabulary (OOV).