Phylogenetic Inference for Language

Nicholas Andrews, Jason Eisner, Mark Dredze

Department of Computer Science, CLSP, HLTCOE
Johns Hopkins University
Baltimore, Maryland 21218

noa@jhu.edu

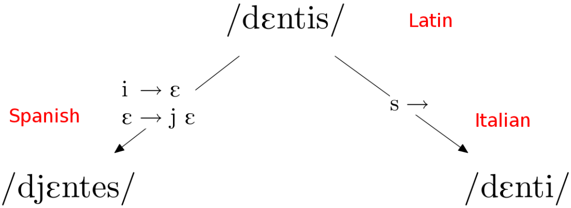
April 23, 2013

Outline

1. Phylogenetic inference?
2. Generative model
3. A sampler sketch
4. Variational EM
5. Experiments

Phylogenetic inference?

Language evolution: e.g. sound change[1]



[1](Bouchard-Côté et al., 2007)

Phylogenetic inference?

Bibliographic entry variation:

Steven Abney, Robert E. Schapire, & Yoram Singer (1999). Boosting applied to tagging and PP attachment. Proc. EMNLP-VLC. New Brunswick, New Jersey: Association for Computational Linguistics

abbreviate names

Abney, S., **Schapire, R**. E., & **Singer, Y.** (1999). Boosting applied to tagging and PP attachment. Proc. EMNLP-VLC. New Brunswick, New Jersey: Association for Computational Linguistics

initials first; shorten to ACL

delete location, shorten venue

S. Abney, **R. E.** Schapire & **Y.** Singer (1999). Boosting applied to tagging and PP attachment. **In** Proc. EMNLP-VLC. New Brunswick, New Jersey. **ACL.**

Abney, S., Schapire, R. E., & Singer, Y. (1999). Boosting applied to tagging and PP attachment. **EMNLP**.

Phylogenetic inference?

Paraphrase:

Papa ate the caviar

substitute "devoured"

add "with a spoon"

Papa devoured the caviar

Papa ate the caviar with a spoon

Active to passive

The caviar was devoured by papa

One Entity, Many Names

Qaddafi, Muammar

Al-Gathafi, Muammar

al-Qadhafi, Muammar

Al Qathafi, Mu’ammar

Al Qathafi, Muammar

El Gaddafi, Moamar

El Kadhafi, Moammar

El Kazzafi, Moamer

معمر محمد عبد السلام أبو منيار القذافي

معمر محمد أبو منيار القذافي

معمر القذافي

أبو محمد

[2]

---

[2] Spence et al, NAACL 2012

Phylogenetic inference?

In each example, there are systematic changes over time:

- **Sound change:** assimilation, metathesis, etc.
- **Bibliographic variation:** typos, abbreviations, punctuation, etc.
- **Paraphrase:** synonyms, voice change, re-arrangements, etc.
- **Name variation:** nicknames, titles, initials, etc.

In each example, there are systematic changes over time:

- **Sound change:** assimilation, metathesis, etc.
- **Bibliographic variation:** typos, abbreviations, punctuation, etc.
- **Paraphrase:** synonyms, voice change, re-arrangements, etc.
- **Name variation:** nicknames, titles, initials, etc.

This talk: **name variation**

1. Phylogenetic inference?
2. Generative model
3. A sampler sketch
4. Variational EM
5. Experiments

A phylogeny is a directed tree rooted at $\diamondsuit$

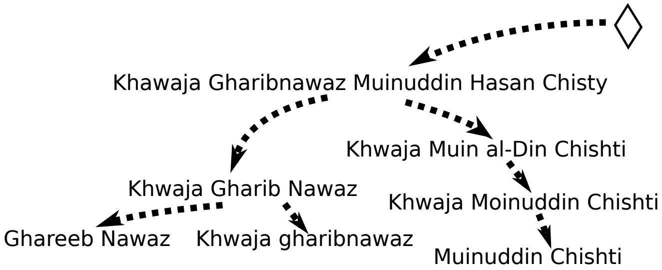


Figure: *A cherry-picked fragment of a phylogeny learned by our model.*

Objects in the model

Names are mentioned in context:

[redacted] @sarah5232 8 Apr
Beliebers held up infinity signs at **Justin**'s **concert** tonight. So beautiful. pic.twitter.com/qwVWrIJctP

Observed?	Description	Example
✓	Name	Justin
	Parent	x_{13}
	Entity	e_{44} (= Justin Bieber)
✓	Type	PERSON
	Topic	6 (= MUSIC)
✓	Document	d_{20}
✓	Language	ENGLISH
✓	Token position	100
	Index	729

Generative model

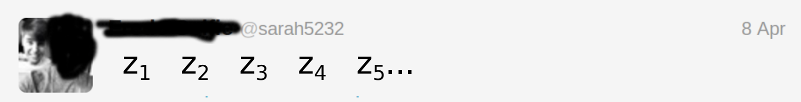
Step 1: Sample a topic z at each position in each document[3] (for all documents in the corpus):


[redacted] @sarah5232 8 Apr

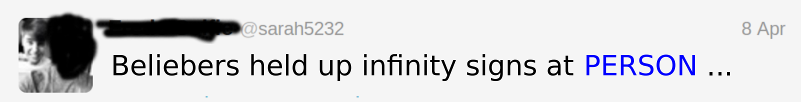
z_1 z_2 z_3 z_4 $z_5 \ldots$

[3] This is just like latent Dirichlet allocation (LDA).

Step 1: Sample a topic z at each position in each document[3] (for all documents in the corpus):



Step 2: Sample either (1) a context word or (2) a named-entity type at each position, conditioned on the topic:



[3] This is just like latent Dirichlet allocation (LDA).

Generative model

Step 3: For the nth named-entity mention y, pick a parent x:

1. Pick $\diamondsuit$ with probability $\frac{\alpha}{n+\alpha}$

$$\begin{tikzcd} \diamondsuit \arrow[d] \\ \text{PERSON}_n \end{tikzcd}$$

Generative model

Step 3: For the nth named-entity mention y, pick a parent x:

1. Pick $\diamondsuit$ with probability $\frac{\alpha}{n+\alpha}$

$$\begin{array}{c} \diamondsuit \\ \downarrow \\ \mathrm{PERSON}_n \end{array}$$

2. Pick a previous mention with probability proportional to $\exp\left(\phi \cdot \mathbf{f}(x, y)\right)$:

$$\begin{array}{c} x \\ \downarrow \\ \mathrm{PERSON}_n \end{array}$$

Features of x and y: topic, entity type, language

Step 4: Generate a name conditioned on the selected parent

1. If the parent is $\diamondsuit$, generate a name from scratch

$$\begin{array}{c}\diamondsuit\\ \downarrow\\ \text{Justin Bieber}\end{array}$$

Step 4: Generate a name conditioned on the selected parent

1. If the parent is $\diamondsuit$, generate a name from scratch

$$\diamondsuit \downarrow \text{Justin Bieber}$$

2. Otherwise:

$$\text{Justin Bieber} \downarrow \text{Justin Bieber}$$

COPY with probability $1 - \mu$

Step 4: Generate a name conditioned on the selected parent

1. If the parent is $\diamondsuit$, generate a name from scratch

$$\diamondsuit \downarrow \text{Justin Bieber}$$

2. Otherwise:

$$\text{Justin Bieber} \downarrow \text{Justin Bieber}$$

$\textsc{copy}$ with probability $1 - \mu$

$$\text{Justin Bieber} \downarrow \text{J.B.}$$

$\textsc{mutate}$ with probability μ

Name variation as mutations

"Mutations" capture different types of name variation:

1. **Transcription errors:** Barack $\rightarrow$ barack
2. **Misspellings:** Barack $\rightarrow$ Barrack
3. **Abbreviations:** Barack Obama $\rightarrow$ Barack O.
4. **Nicknames:** Barack $\rightarrow$ Barry
5. **Dropping words:** Barack Obama $\rightarrow$ Barack

Mutation via probabilistic finite-state transducers

The mutation model is a **probabilistic finite-state transducer** with four character operations: COPY, SUBSTITUTE, DELETE, INSERT

- Character operations are conditioned on the right input character
- Latent regions of contiguous edits
- Back-off smoothing

Transducer parameters $\boldsymbol{\theta}$ determine the probability of being in different regions, and of the different character operations

Example: Mutating a name

Mr. Robert Kennedy

Mr. Bobby Kennedy

Example mutation

```
M r . _ R o b e r t     _ K e n n e d y $
M r . _[
```

Beginning of edit region

Example: Mutating a name

Mr. Robert Kennedy

Mr. Bobby Kennedy

Example mutation

```
M r . _ R o b e r t     _ K e n n e d y $
M r . _ [B
```

1 substitution operation: (R, B)

Example: Mutating a name

Mr. Robert Kennedy

$\downarrow$

Mr. Bobby Kennedy

Example mutation

```
M r . _ R o b e r t       _ K e n n e d y $
M r . _[B o b
```

2 copy operations: (ε, o), (ε, b)

Example: Mutating a name

Mr. Robert Kennedy

↓

Mr. Bobby Kennedy

Example mutation

```
M r . _ R o b e r t     _ K e n n e d y $
M r . _[B o b
```

3 deletion operations: (e,ε), (r,ε), (t, ε)

Example: Mutating a name

Mr. Robert Kennedy

$\downarrow$

Mr. Bobby Kennedy

Example mutation

```
M r . _ R o b e r t     _ K e n n e d y$
M r . _[B o b         b y
```

2 insertion operations: (ε,b), (ε,y)

Example: Mutating a name

Mr. Robert Kennedy

↓

Mr. Bobby Kennedy

Example mutation

```
M r . _ R o b e r t   _ K e n n e d y $
M r . _[B o b      b y]
```

End of edit region

Example: Mutating a name

Mr. Robert Kennedy

Mr. Bobby Kennedy

Example mutation

```
M r . _ R o b e r t     _ K e n n e d y $
M r . _[B o b       b y]_ K e n n e d y $
```

1. Phylogenetic inference?
2. Generative model
3. A sampler sketch
4. Variational EM
5. Experiments

The latent variables in the model are[4]

- The spanning tree over tokens $\mathbf{p}$
- The token permutation $\mathbf{i}$
- The topics of all named-entity and context tokens $\mathbf{z}$

Inference requires marginalizing over the latent variables:

$$\mathrm{Pr}_{\phi,\theta}(\mathbf{x}) = \sum_{\mathbf{p},\mathbf{i},\mathbf{z}} \mathrm{Pr}_{\phi,\theta}(\mathbf{x}, \mathbf{z}, \mathbf{i}, \mathbf{p})$$

[4] The mutation model also has latent alignments

Inference

The latent variables in the model are

- The spanning tree over tokens $\mathbf{p}$
- The token permutation $\mathbf{i}$
- The topics of all named-entity and context tokens $\mathbf{z}$

Inference requires marginalizing over the latent variables:

$$\Pr_{\phi,\theta}(\mathbf{x}) = \sum_{\mathbf{p},\mathbf{i},\mathbf{z}} \Pr_{\phi,\theta}(\mathbf{x}, \mathbf{z}, \mathbf{i}, \mathbf{p})$$

This sum is intractable to compute ☹

Inference

The latent variables in the model are

- The spanning tree over tokens $\mathbf{p}$
- The token permutation $\mathbf{i}$
- The topics of all named-entity and context tokens $\mathbf{z}$

Inference requires marginalizing over the latent variables:

$$
\begin{aligned}
\Pr_{\phi,\theta}(\mathbf{x}) &= \cancel{\sum_{\mathbf{p},\mathbf{i},\mathbf{z}} \Pr_{\phi,\theta}(\mathbf{x},\mathbf{z},\mathbf{i},\mathbf{p})} \\
&\approx \frac{1}{N}\sum_{n=1}^{N} \Pr_{\phi,\theta}(\mathbf{x},\mathbf{z}_n,\mathbf{i}_n,\mathbf{p}_n)
\end{aligned}
$$

But we can sample from the posterior! ☺

Key idea: sampling $(\mathbf{p}, \mathbf{i}, \mathbf{z})$ jointly is hard, but sampling from the conditional for each variable is easy(ier)

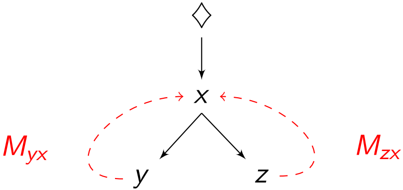
Key idea: sampling $(\mathbf{p}, \mathbf{i}, \mathbf{z})$ jointly is hard, but sampling from the conditional for each variable is easy(ier)

Procedure:

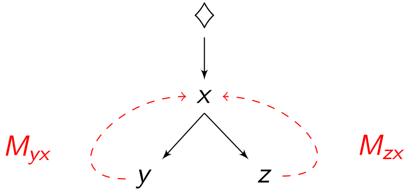
- Initialize $(\mathbf{p}, \mathbf{i}, \mathbf{z})$.
- For $n = 1$ to N:
  1. Resample a permutation $\mathbf{i}$ given all other variables.
  2. Resample the topic vector $\mathbf{z}$, similarly.
  3. Resample the phylogeny $\mathbf{p}$, similarly.
  4. Output the current sample $(\mathbf{p}, \mathbf{i}, \mathbf{z})$.

Steps 1 and 2 are Metropolis-Hastings proposals

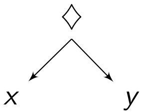
Step 1: Run belief propagation with messages M_{ij} directed from the leaves to the root $\diamondsuit$



Step 1: Run belief propagation with messages M_{ij} directed from the leaves to the root $\diamondsuit$



Step 2: Sample topics $\mathbf{z}$ from $\diamondsuit$ downwards proportional to the belief at each vertex, conditioned on previously sampled topics



(a) Compatible with both (x, y) and (y, x).

(b) Compatible with a single permutation: (x, y).

Each edge between non-root vertices yields a constraint on possible permutations:

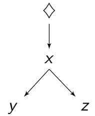
Example

$$\begin{tikzcd}
 & \Diamond \arrow[d] & \\
 & x \arrow[dl] \arrow[dr] & \\
y & & z
\end{tikzcd}$$

yields two constraints: $x \prec y$ and $x \prec z$.

Each edge between non-root vertices yields a constraint on possible permutations:

Example



yields two constraints: $x \prec y$ and $x \prec z$.

Sampling uniformly from the set of permutations respecting these constraints is a simple recursive procedure:

```
def unif_perm(u):
  yield u
  for x in unif_shuffle([ unif_perm(x) for x in children[u] ]):
    yield x
```

Conditioned on topics and a permutation of the tokens, sample a parent x for each mention y with probability:

$$\propto \underbrace{\Pr_{\phi}(x, y)}_{\text{affinity model}} \cdot \underbrace{\Pr_{\boldsymbol{\theta}}(x.n, y.n)}_{\text{transducer model}}$$

No cycles, since the mention permutation $\mathbf{i}$ is known.

Outline

1 Phylogenetic inference?

2 Generative model

3 A sampler sketch

4 Variational EM

5 Experiments

The sampler is still running ☹

The sampler is still running ☹

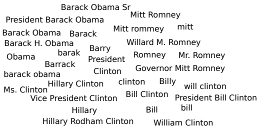
We report experiments from our EMNLP 2012 paper + followup experiments, which use a simpler model:

- No context/topics: only the transducer parameters $\boldsymbol{\theta}$ need to be estimated
- Type-level inference and supervision: vertices in the phylogeny represent distinct name types rather than name tokens

Inference

Input: An unaligned corpus of names ("bag-of-words")

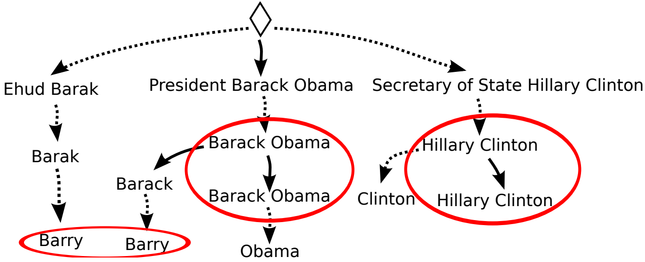
- The order in which the tokens were generated is unknown
- No "inputs" or "outputs" are known for the mutation model



Output: A distribution over name phylogenies parametrized by transducer parameters $\boldsymbol{\theta}$

Type phylogeny vs token phylogeny

The generative model is over **tokens** (name mentions)

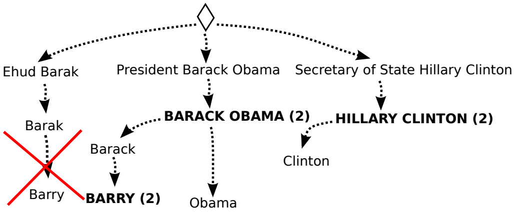


But we do **type-level** inference for the following reasons:

1. Allows faster inference
2. Allows type-level supervision

Type phylogeny vs token phylogeny

We collapse all COPY edges into a single vertex



- The first token in each collapsed vertex is a MUTATION, and the rest are COPIES
- Every edge in the phylogeny now corresponds to a mutation
- Approximation: disallow multiple tokens of the same type to be derived from mutations

Edge weights

- New names: edges from $\diamondsuit$ to a name x:

$$\delta(x \mid \diamondsuit) = \alpha \cdot p(x \mid \diamondsuit)$$

- Mutations: edges from a name x to a name y:

$$\delta(y \mid x) = \mu \cdot p(y \mid x) \cdot \frac{n_x}{n_y + 1}$$

Approximation: Edges weights are not *quite* edge factored. We are making an approximation of the form

$$\mathbb{E} \prod_y \delta(y \mid \mathrm{pa}(y)) \approx \prod_y \mathbb{E}\delta(y \mid \mathrm{pa})$$

Inference via EM

Iterate until convergence:

1. **E-step:** Given $\boldsymbol{\theta}$, compute a *distribution* over name phylogenies
2. **M-step:** Re-estimate transducer parameters $\boldsymbol{\theta}$ given marginal edge probabilities.
   - This step sums over alignments for each (x, y) string pair using forward-backward
   - Each (x, y) pair may be viewed as a training example weighted by the marginal probability of the edge from x to y

E-step: marginalizing over latent variables

The latent variables in the model are:

1. Name phylogeny (spanning tree) relating names as inputs and/or outputs
2. Character alignments from potential input names x to output names y

We use the Matrix-Tree theorem for directed graphs (Tutte, 1984) to efficiently evaluate marginal probabilities:

1. Partition function (sum over phylogenies)
2. Edge marginals

1. Phylogenetic inference?
2. Generative model
3. A sampler sketch
4. Variational EM
5. Experiments

- We collected a corpus of **Wikipedia redirect strings** used as examples of names variations
  - Filtered down to a subset 77489 people from English Wikipedia (Examples in the next slide!)
- The frequency of each variation is estimated using the **Google crosswiki dataset**[5]
  - Dictionary of anchor strings linking to English Wikipedia articles
  - Collected "by crawling a reasonably large approximation of the entire web"

[5] Spitkovsky and Chang, 2012

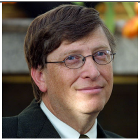
Example Wikipedia redirects

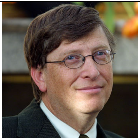
Ho Chi Minh
Ho chi mihn
Ho-Chi Minh
Ho Chih-minh

Example Wikipedia redirects

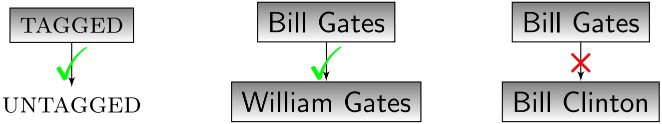
	Ho Chi Minh Ho chi mihn Ho-Chi Minh Ho Chih-minh
	Guy Fawkes Guy fawkes Guy faux Guy foxe

Example Wikipedia redirects

	Ho Chi Minh Ho chi mihn Ho-Chi Minh Ho Chih-minh
	Guy Fawkes Guy fawkes Guy faux Guy foxe
	Bill Gates Lord Billy William Gates III William H. Gates

	Ho Chi Minh Ho chi mihn Ho-Chi Minh Ho Chih-minh
	Guy Fawkes Guy fawkes Guy faux Guy foxe
	Bill Gates Lord Billy William Gates III William H. Gates
	Billll Clinton William J. Blythe IV William Clinton President Clinton

Type-level supervision is incorporated by tagging vertices with unique IDs and enforcing that they agree from parent to child:



Experiment 1: Evaluating the transducer

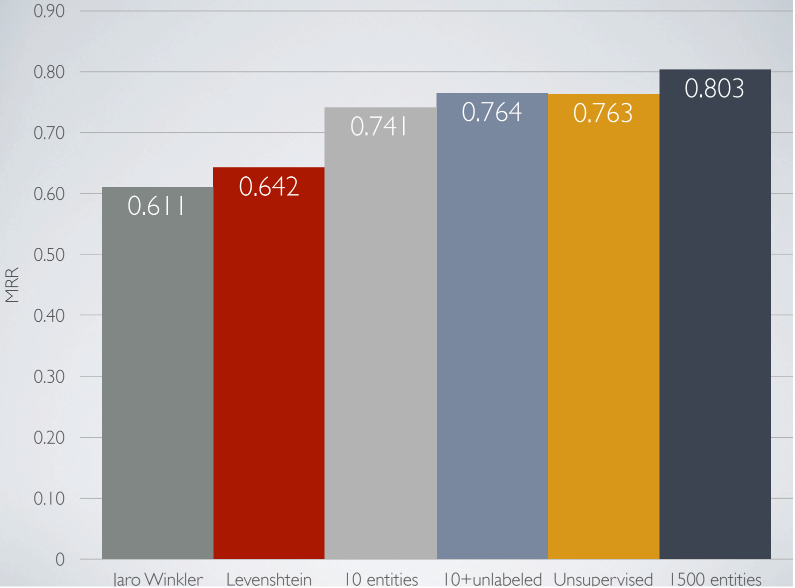
Procedure:

- At train time:
  1. Estimate the transducer parameters $\boldsymbol{\theta}$
- At test time:
  1. For each name x in the test set, rank all other names y by the transducer probability

$$\Pr_{\boldsymbol{\theta}}(y \mid x)$$

  2. Compute the mean reciprocal rank (MRR) over all names

Experiment 1: Evaluating the transducer



Experiment 2: Evaluating the phylogeny

Step 1: Estimate θ via EM on the training corpus
Step 2: Find the highest scoring tree [6]

President Bill Clinton

Guido Fawkes

Bill Gates

Guy Fawkes

Lord Billy

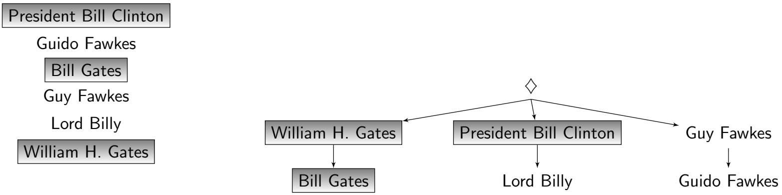
William H. Gates

Input: "bag of words."

[6] $O(m \log n)$ for graphs of n vertices and m edges

Experiment 2: Evaluating the phylogeny

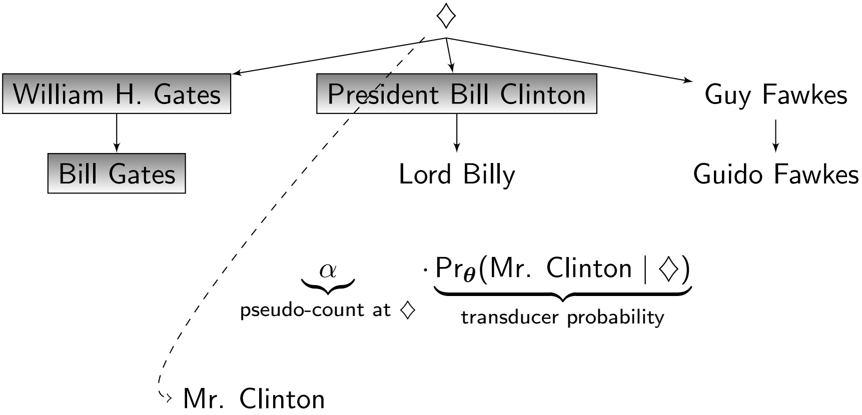
Step 1: Estimate θ via EM on the training corpus
Step 2: Find the highest scoring tree [6]



Input: "bag of words." **Output:** 1-best tree

[6] $O(m \log n)$ for graphs of n vertices and m edges

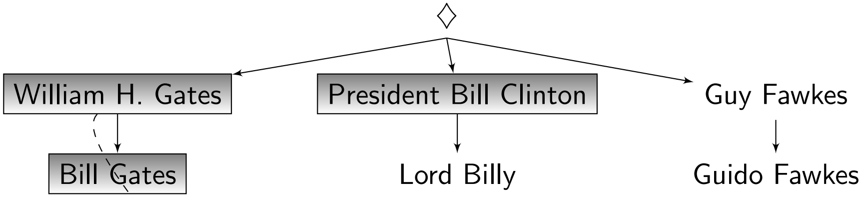
Step 3: Attach each name in the test corpus to its most likely parent in the 1-best tree



$$\underbrace{\alpha}_{\text{pseudo-count at } \diamondsuit} \cdot \underbrace{\Pr_{\boldsymbol{\theta}}(\text{Mr. Clinton} \mid \diamondsuit)}_{\text{transducer probability}}$$

Experiment 2: Evaluating the phylogeny

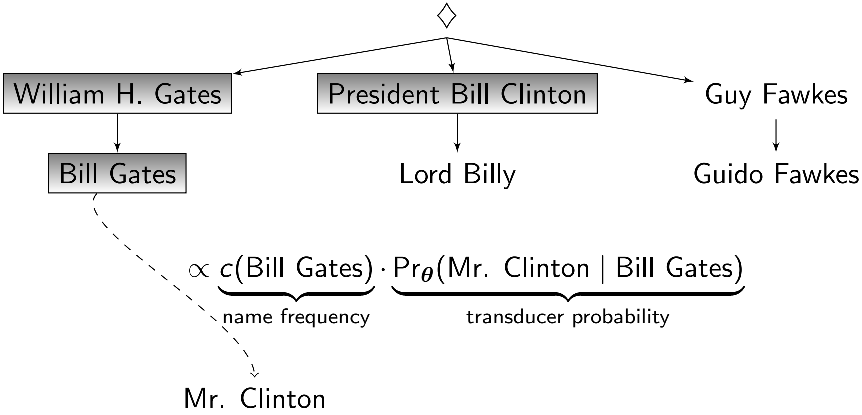
Step 3: Attach each name in the test corpus to its most likely parent in the 1-best tree



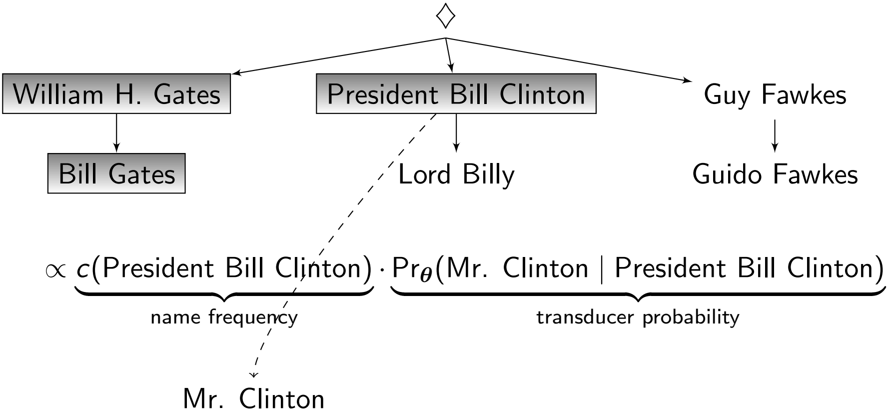
$$\propto \underbrace{c(\text{William H. Gates})}_{\text{name frequency}} \cdot \underbrace{\Pr_{\boldsymbol{\theta}}(\text{Mr. Clinton} \mid \text{William H. Gates})}_{\text{transducer probability}}$$

Mr. Clinton

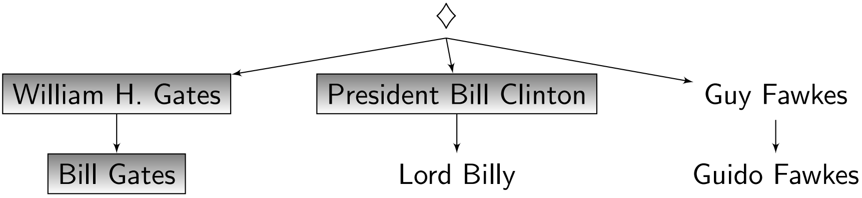
Step 3: Attach each name in the test corpus to its most likely parent in the 1-best tree



Step 3: Attach each name in the test corpus to its most likely parent in the 1-best tree



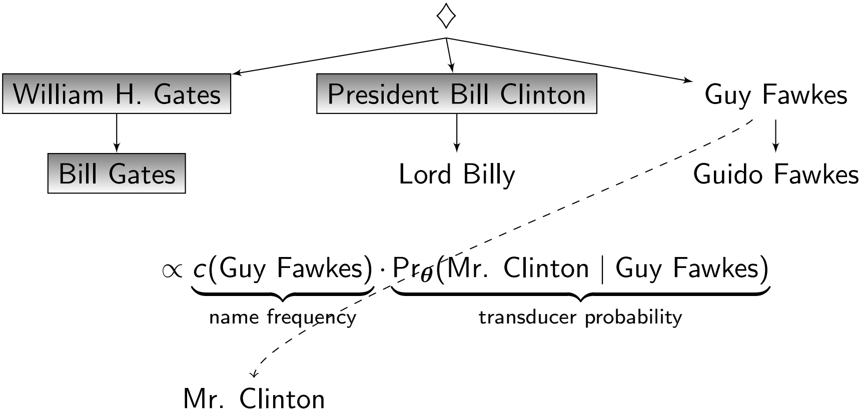
Step 3: Attach each name in the test corpus to its most likely parent in the 1-best tree



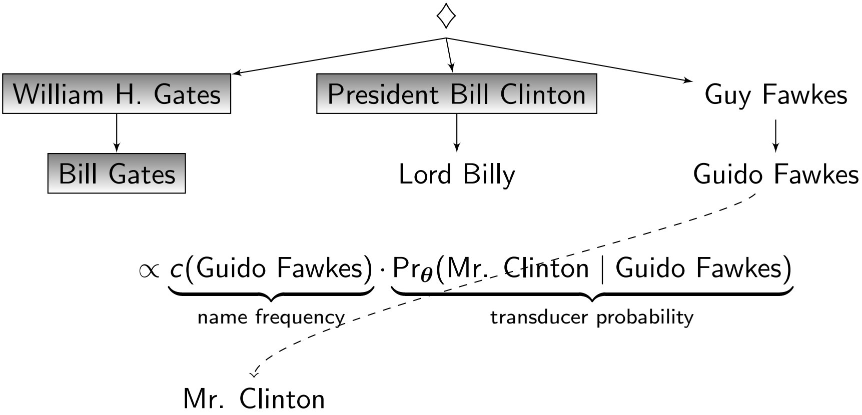
$$\propto \underbrace{c(\text{Lord Billy})}_{\text{name frequency}} \cdot \underbrace{\Pr_{\boldsymbol{\theta}}(\text{Mr. Clinton} \mid \text{Lord Billy})}_{\text{transducer probability}}$$

Mr. Clinton

Step 3: Attach each name in the test corpus to its most likely parent in the 1-best tree



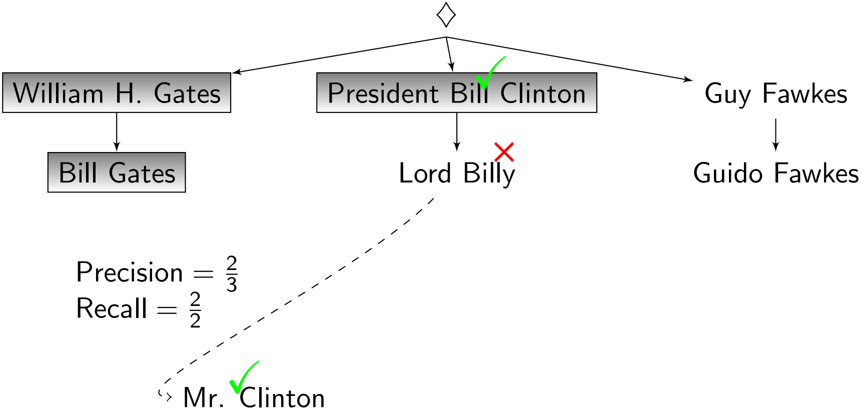
Step 3: Attach each name in the test corpus to its most likely parent in the 1-best tree



$$\propto \underbrace{c(\text{Guido Fawkes})}_{\text{name frequency}} \cdot \underbrace{\Pr_{\boldsymbol{\theta}}(\text{Mr. Clinton} \mid \text{Guido Fawkes})}_{\text{transducer probability}}$$

Experiment 2: Evaluating the phylogeny

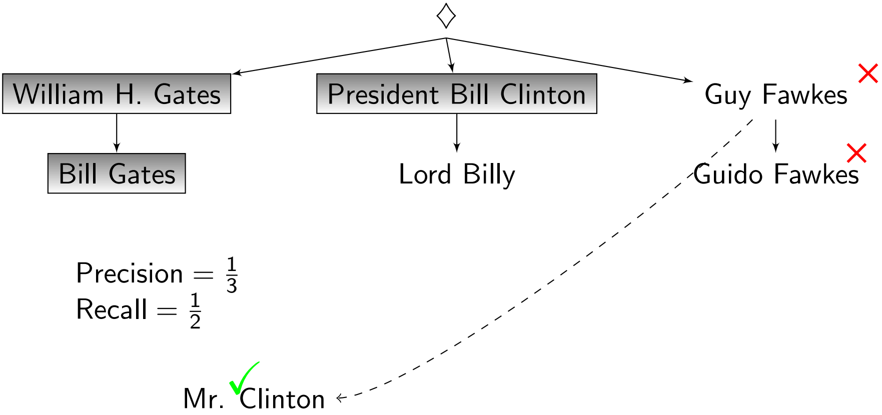
Step 4: Calculate macro-averaged precision and recall for each test name



$$\text{Precision} = \frac{2}{3}$$
$$\text{Recall} = \frac{2}{2}$$

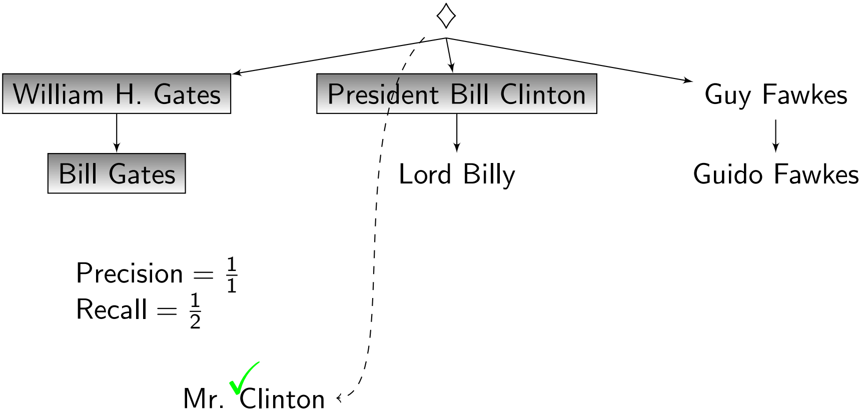
Experiment 2: Evaluating the phylogeny

Step 4: Calculate macro-averaged precision and recall for each test name



Precision $= \frac{1}{3}$
Recall $= \frac{1}{2}$

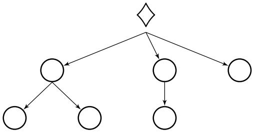
Step 4: Calculate macro-averaged precision and recall for each test name



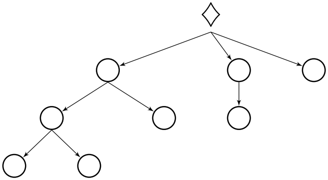
Precision $= \frac{1}{1}$
Recall $= \frac{1}{2}$

We compare to two baselines:

1. Flat tree



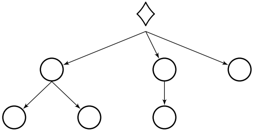
Flat tree: depth ≤ 2



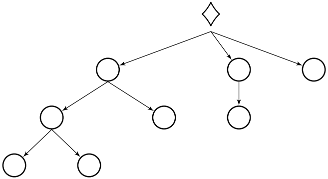
Unrestricted tree

We compare to two baselines:

1. Flat tree



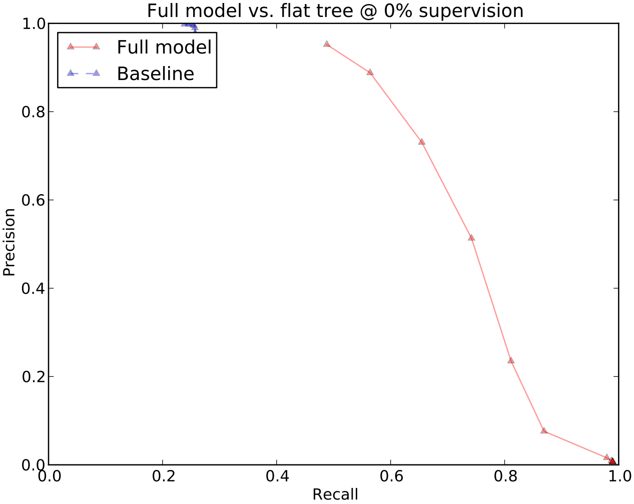
Flat tree: depth ≤ 2



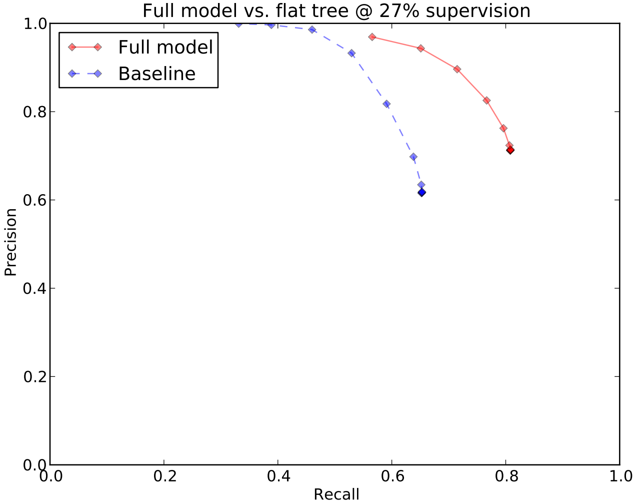
Unrestricted tree

2. Weak transducer
   - No latent edit regions
   - Only 3 degrees of freedom: the weights of different edit operations

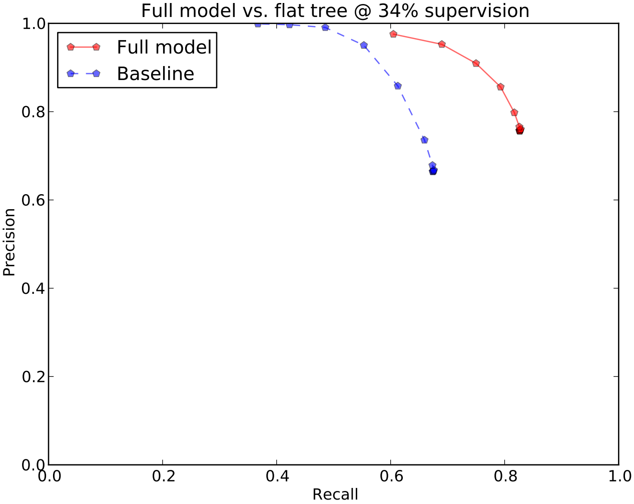
Comparison to flat tree



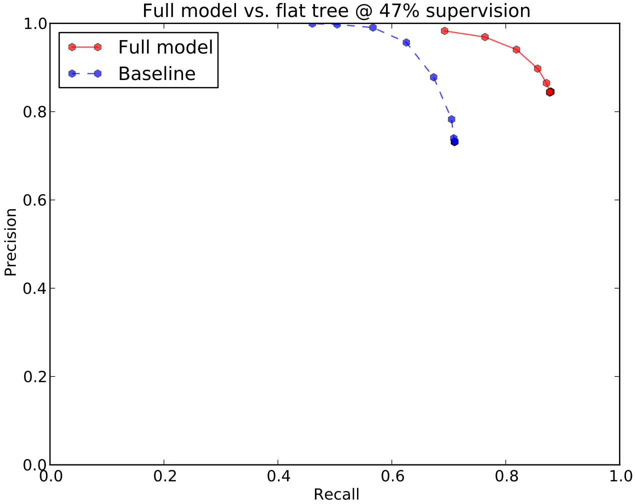
Comparison to flat tree



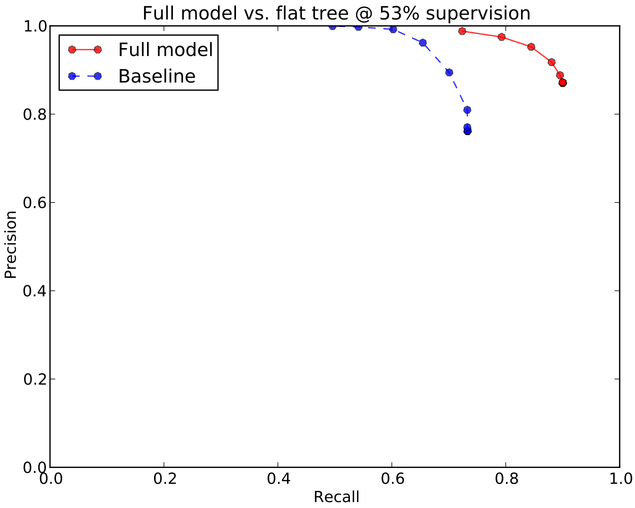
Comparison to flat tree



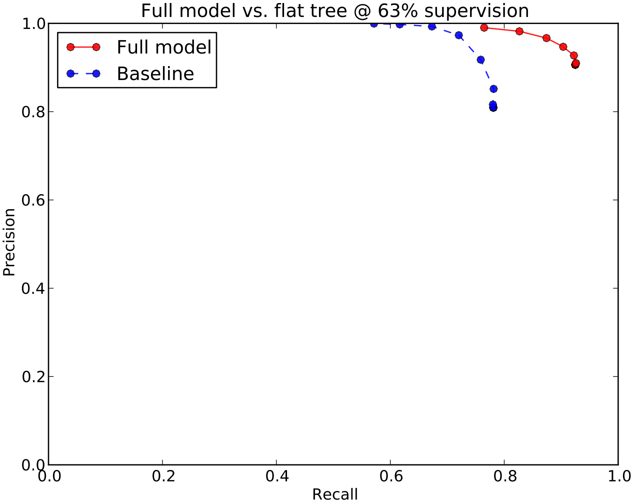
Comparison to flat tree



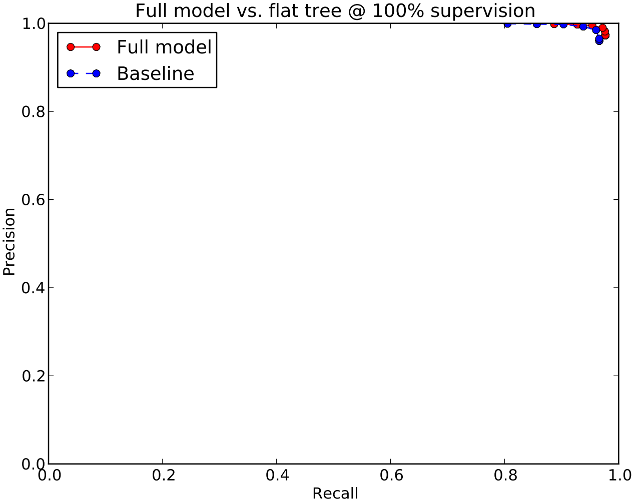
Comparison to flat tree



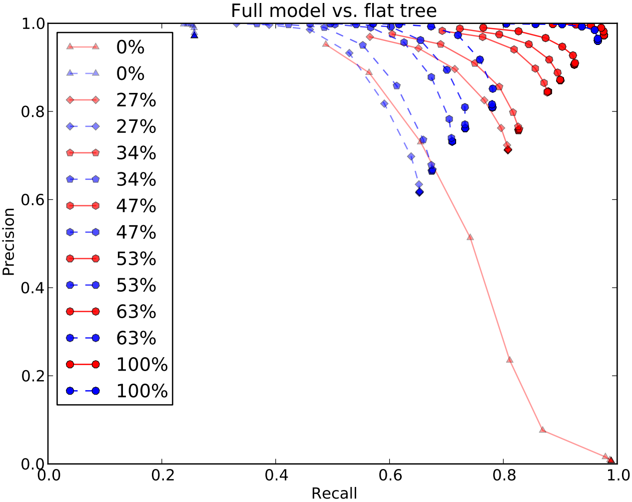
Comparison to flat tree



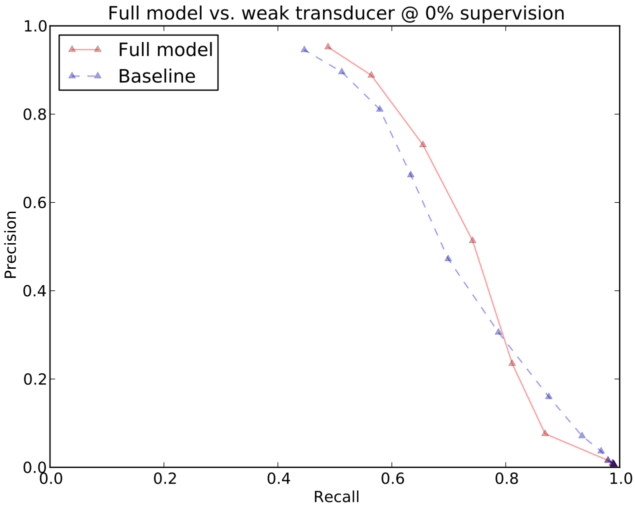
Comparison to flat tree



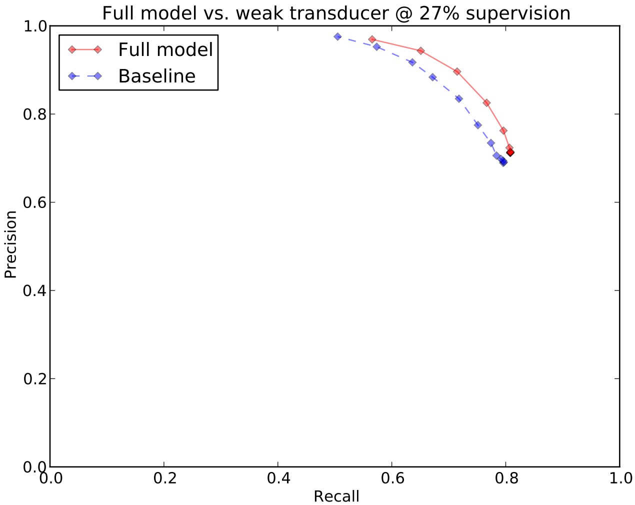
Comparison to flat tree



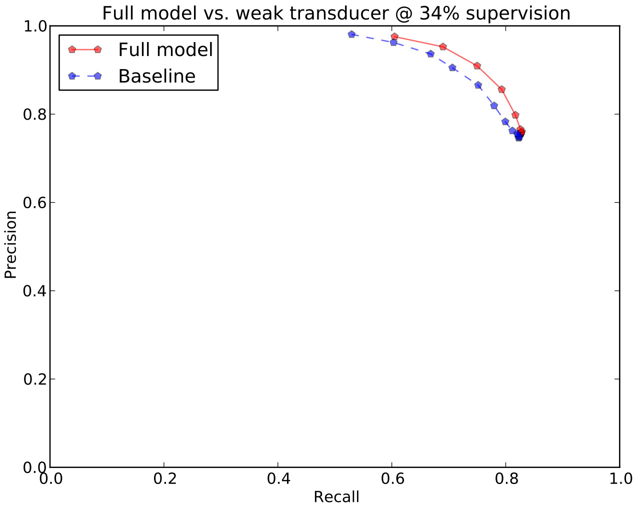
Comparison to weak transducer



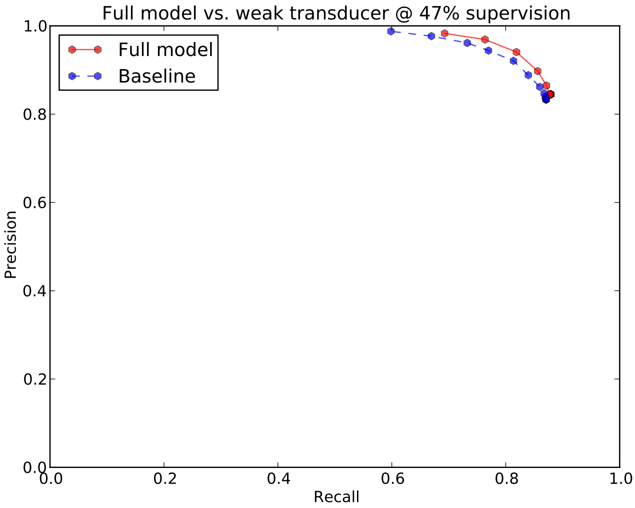
Comparison to weak transducer



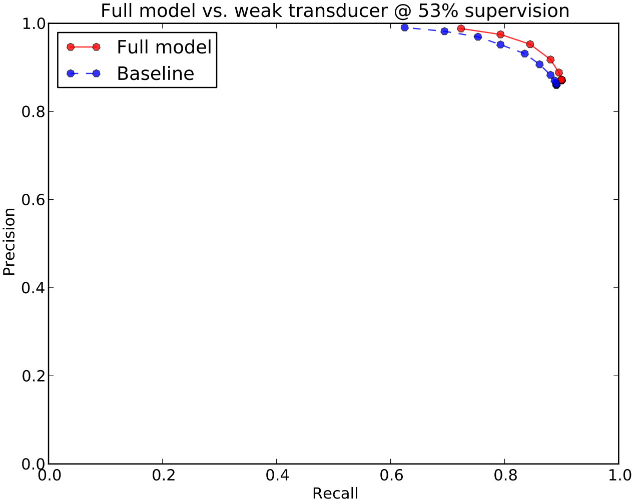
Comparison to weak transducer



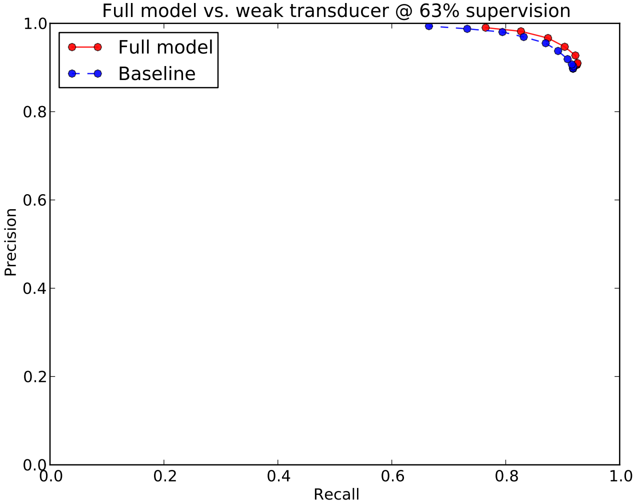
Comparison to weak transducer



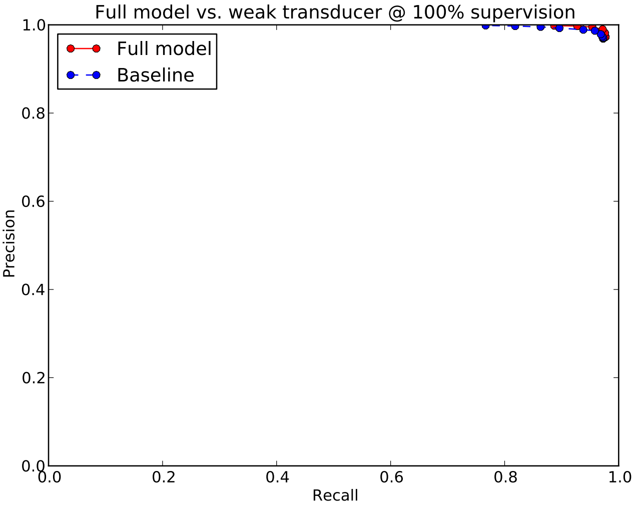
Comparison to weak transducer



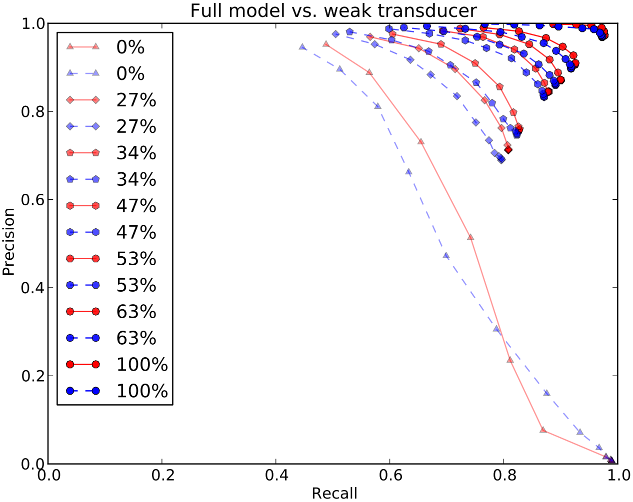
Comparison to weak transducer



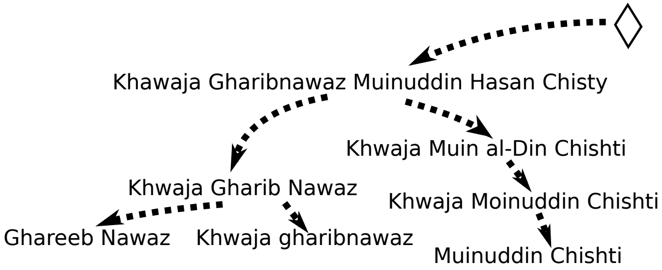
Comparison to weak transducer



Comparison to weak transducer



The End



Thanks! Questions?