Structure Compilation:
Trading Structure for Features

ICML 2008 Helsinki, Finland

July 8, 2008

Percy Liang
UC Berkeley

Hal Daumé
U of Utah

Dan Klein
UC Berkeley

Structured prediction:

Part-of-speech tagging (POS)

Structured prediction:

Part-of-speech tagging (POS)

Named-entity recognition (NER)

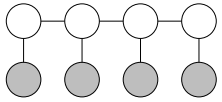
Structured prediction:

Part-of-speech tagging (POS)

Named-entity recognition (NER)

Current methods:

Structured models: accurate but slow



conditional random fields (CRFs) with loopy graphs, large tag sets

Structured prediction:

y: DT — NNP — NNP — VBD

x: The European Commision agreed

Part-of-speech tagging (POS)

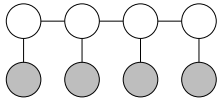
y: O — B-ORG — I-ORG — O

x: The European Commission agreed

Named-entity recognition (NER)

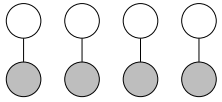
Current methods:

Structured models: accurate but slow



conditional random fields (CRFs) with loopy graphs, large tag sets

Independent models: less accurate but fast



independent logistic regressions (ILRs)

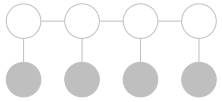
Structured prediction:

Part-of-speech tagging (POS)

Named-entity recognition (NER)

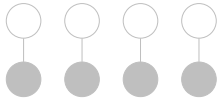
Current methods:

Structured models: accurate but slow



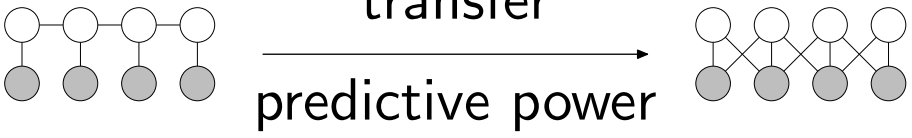
conditional random fields (CRFs) with loopy graphs, large tag sets

Independent models: less accurate but fast



independent logistic regressions (ILRs)

Our goal:



accurate and fast **at test time**

Structured prediction:

y: DT — NNP — NNP — VBD

x: The European Commision agreed

Part-of-speech tagging (POS)

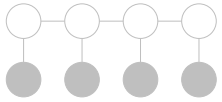
y: O — B-ORG — I-ORG — O

x: The European Commission agreed

Named-entity recognition (NER)

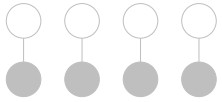
Current methods:

Structured models: accurate but slow



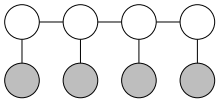
conditional random fields (CRFs) with loopy graphs, large tag sets

Independent models: less accurate but fast



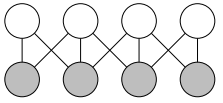
independent logistic regressions (ILRs)

Our goal:



transfer

predictive power



accurate and fast **at test time**

Questions: are independent models...

- ...expressive enough (approximation error)?

Structured prediction:

y:	DT	NNP	NNP	VBD
x:	The	European	Commision	agreed

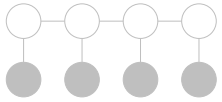
Part-of-speech tagging (POS)

y:	O	B-ORG	I-ORG	O
x:	The	European	Commission	agreed

Named-entity recognition (NER)

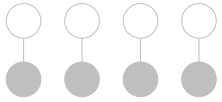
Current methods:

Structured models: accurate but slow



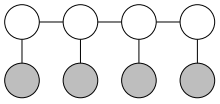
conditional random fields (CRFs) with loopy graphs, large tag sets

Independent models: less accurate but fast



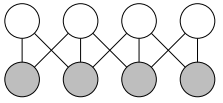
independent logistic regressions (ILRs)

Our goal:



transfer

predictive power

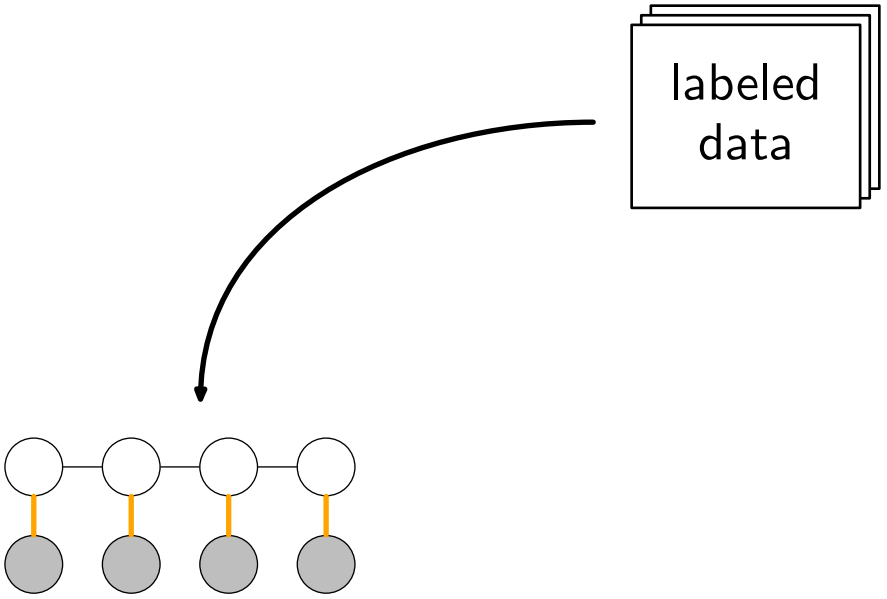


accurate and fast **at test time**

Questions: are independent models...

- ...expressive enough (approximation error)?
- ...easy to learn (estimation error)?

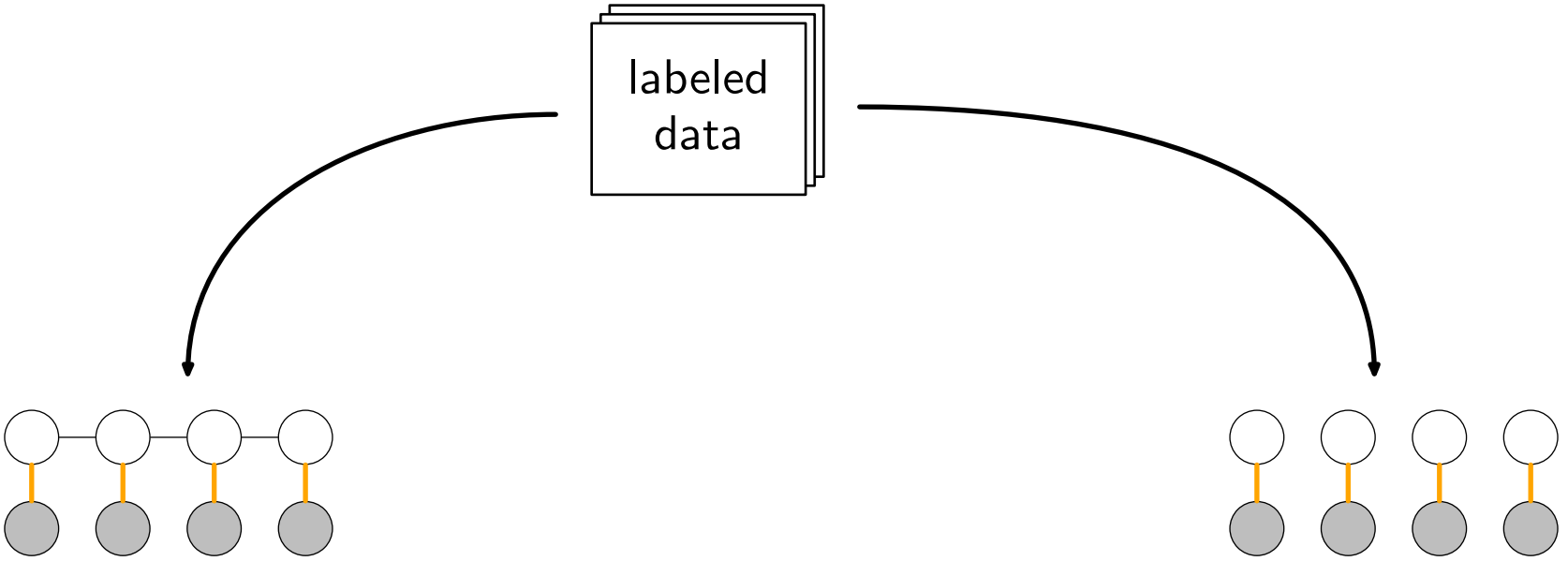
Some empirical motivation



CRF(f_1) POS: 95.0%
NER: 75.3%

f_1: words/prefixes/suffixes/forms

Some empirical motivation

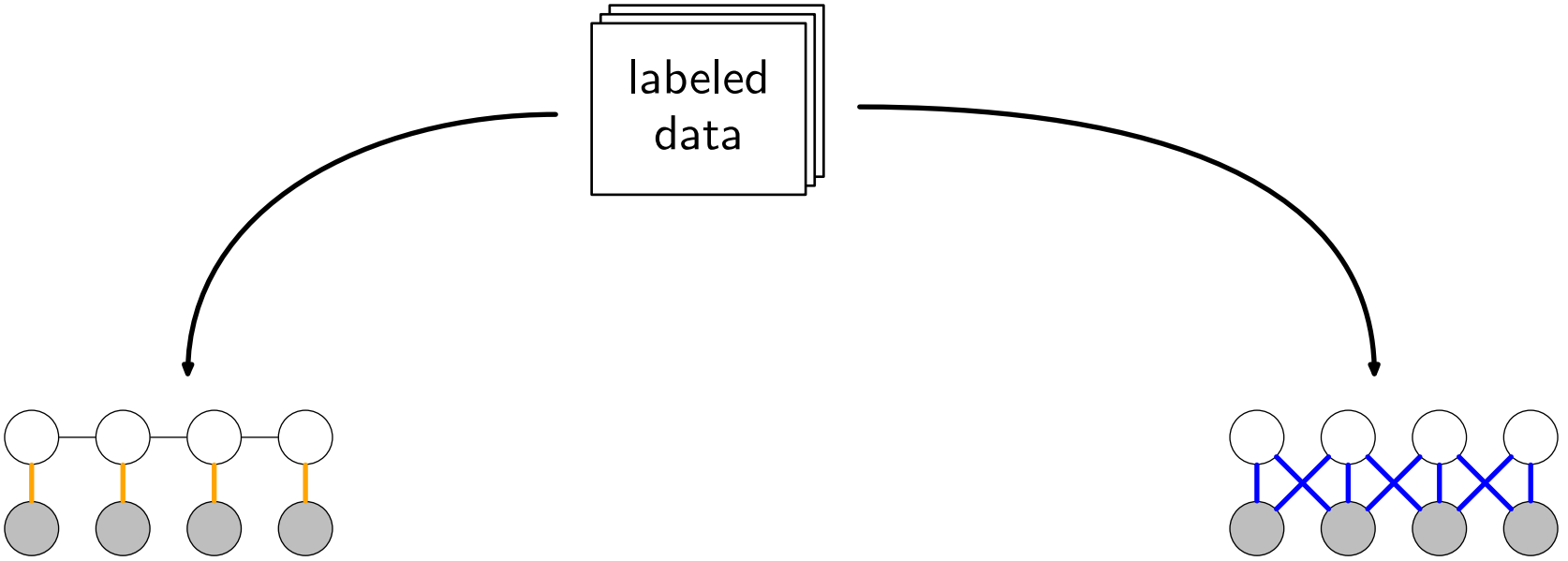


CRF(f_1) POS: 95.0%
NER: 75.3%

ILR(f_1) POS: 91.7%
NER: 69.1%

f_1: words/prefixes/suffixes/forms

Some empirical motivation



CRF(f_1) POS: 95.0%
NER: 75.3%

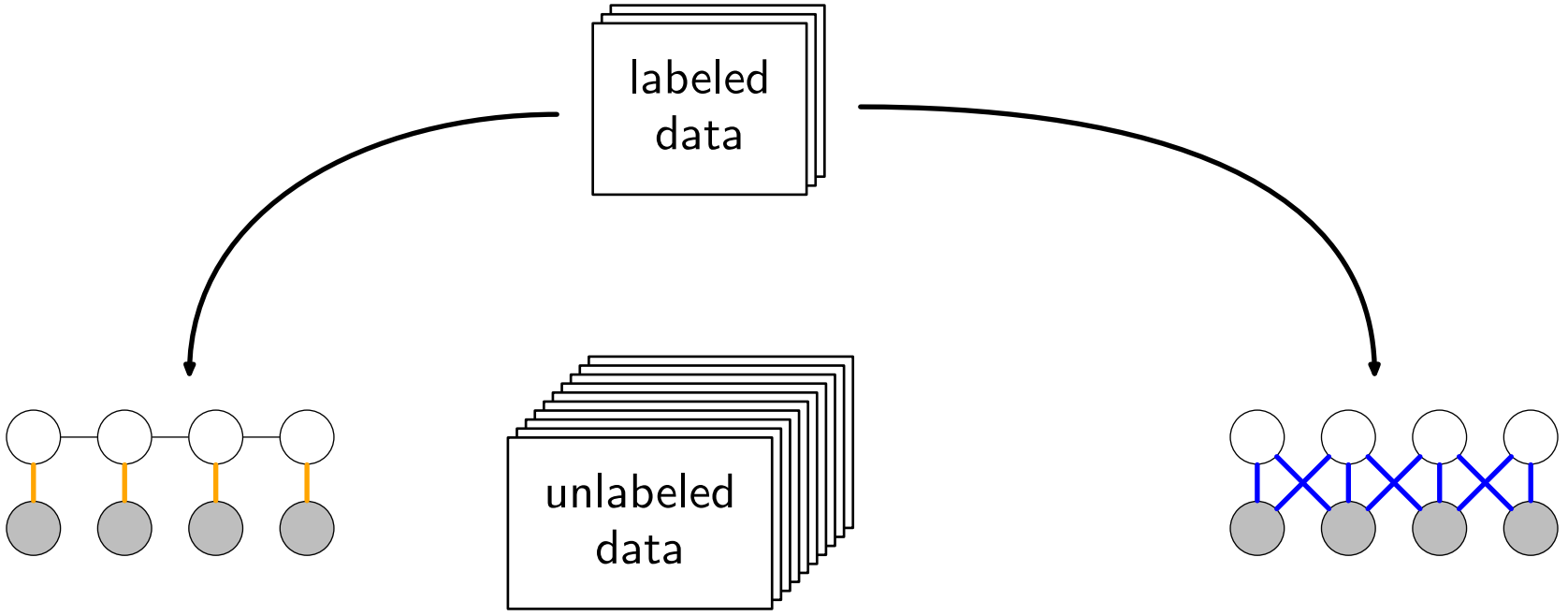
ILR(f_1) POS: 91.7%
NER: 69.1%

ILR(f_2) POS: 94.4%
NER: 66.2%

f_1: words/prefixes/suffixes/forms

f_2: f_1 applied to larger radius

Some empirical motivation



CRF(f_1) POS: 95.0%
NER: 75.3%

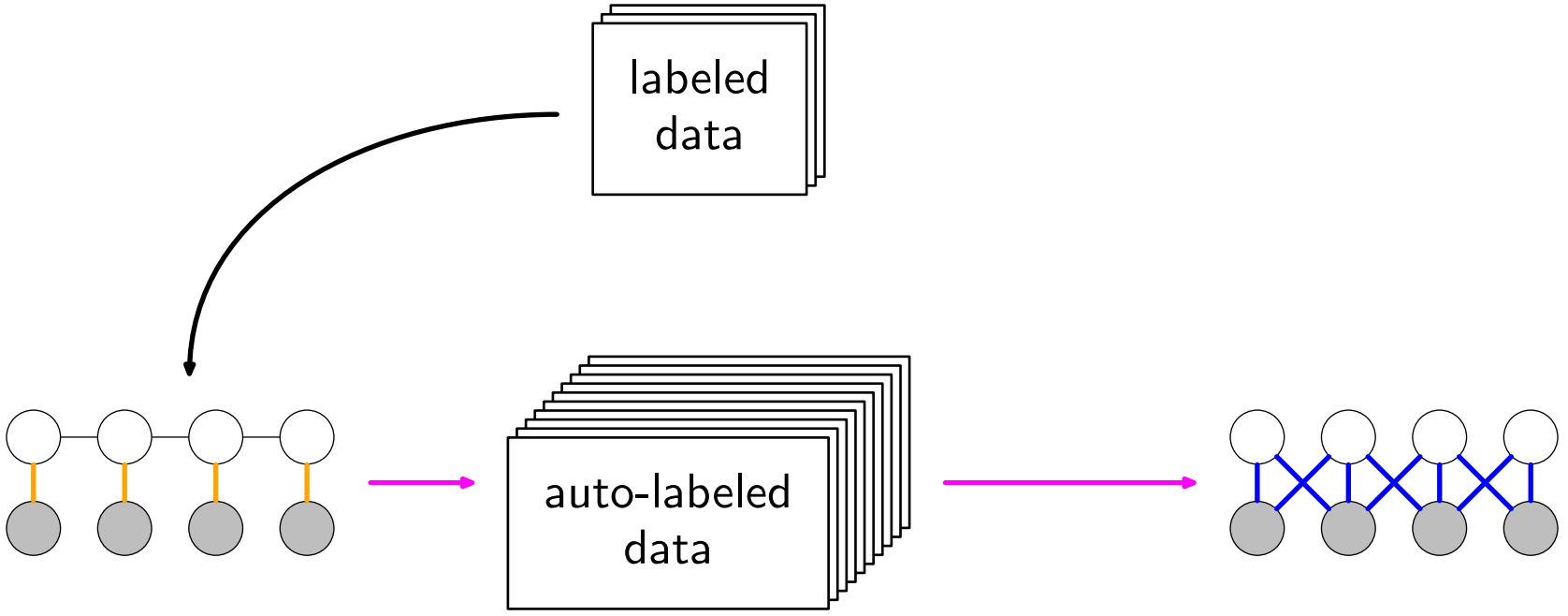
ILR(f_1) POS: 91.7%
NER: 69.1%

ILR(f_2) POS: 94.4%
NER: 66.2%

f_1: words/prefixes/suffixes/forms

f_2: f_1 applied to larger radius

Some empirical motivation



CRF(f_1) POS: 95.0%
NER: 75.3%

f_1: words/prefixes/suffixes/forms

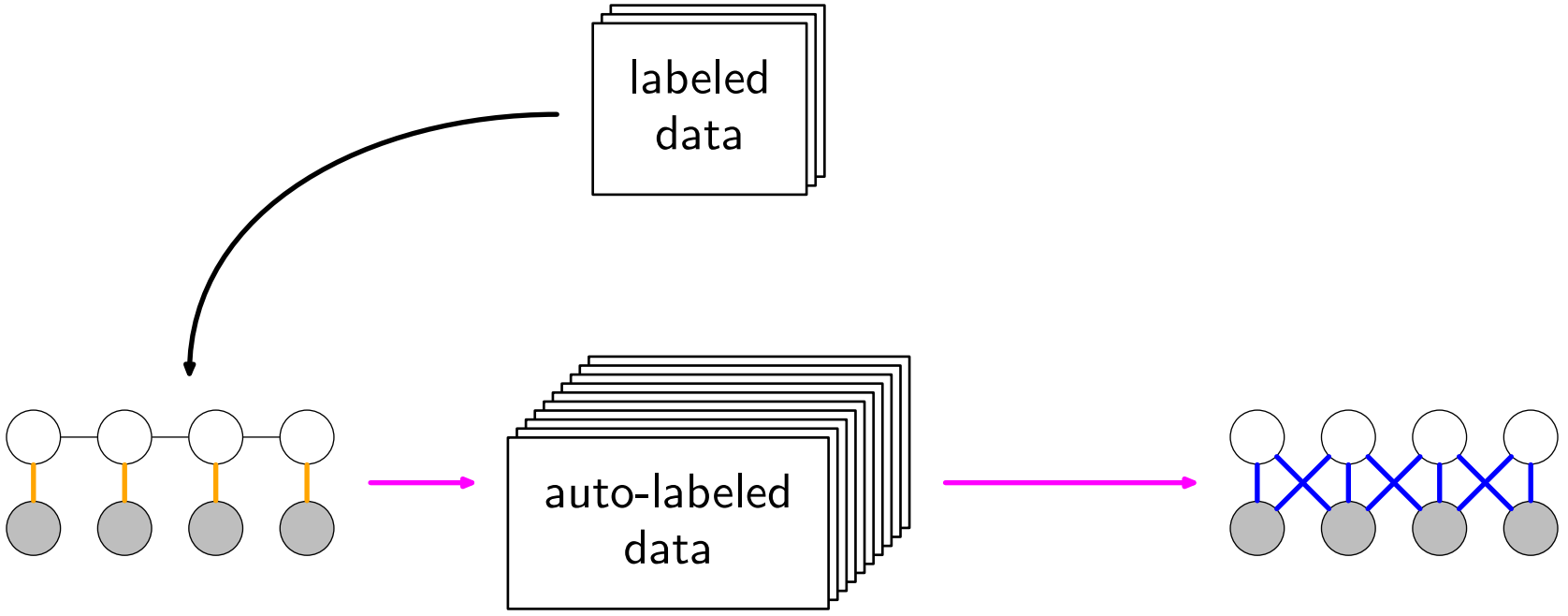
f_2: f_1 applied to larger radius

ILR(f_1) POS: 91.7%
NER: 69.1%

ILR(f_2) POS: 94.4%
NER: 66.2%

ILR(f_2) [compiled] POS: 95.0%
NER: 72.7%

Some empirical motivation



CRF(f_1) POS: 95.0% NER: 75.3%

ILR(f_1) POS: 91.7% NER: 69.1%

ILR(f_2) POS: 94.4% NER: 66.2%

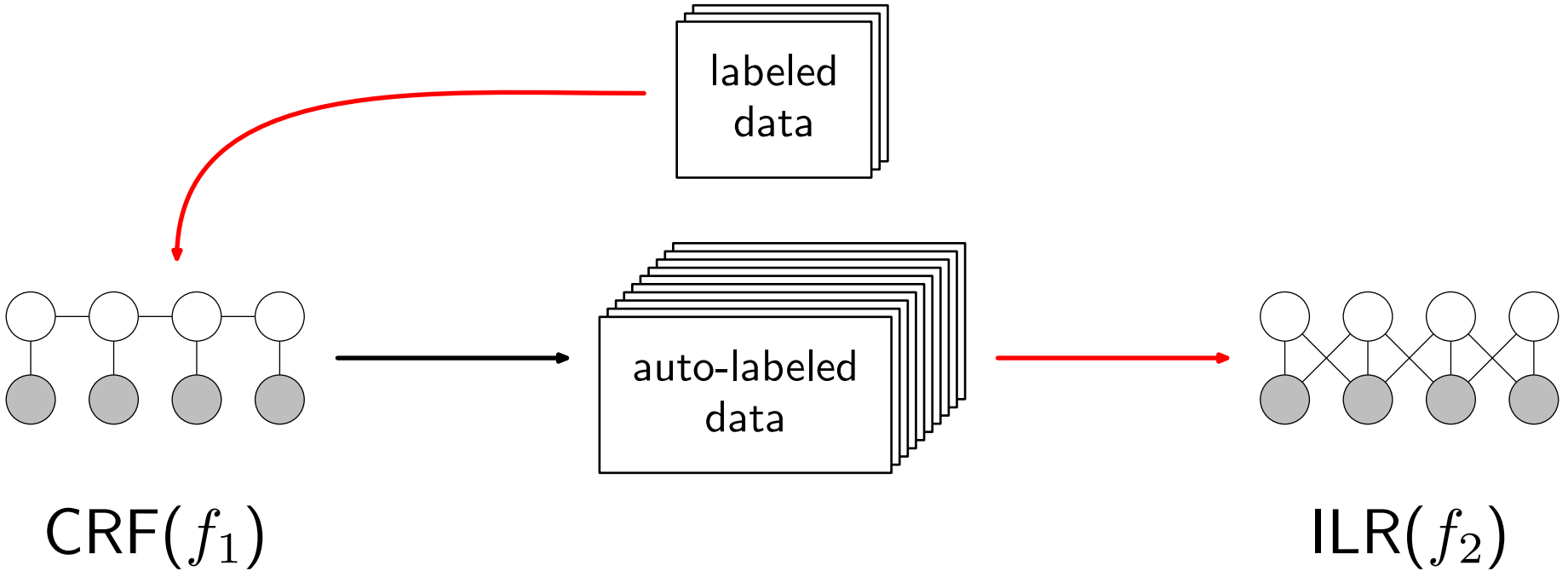
ILR(f_2) [compiled] POS: 95.0% NER: 72.7%

f_1: words/prefixes/suffixes/forms

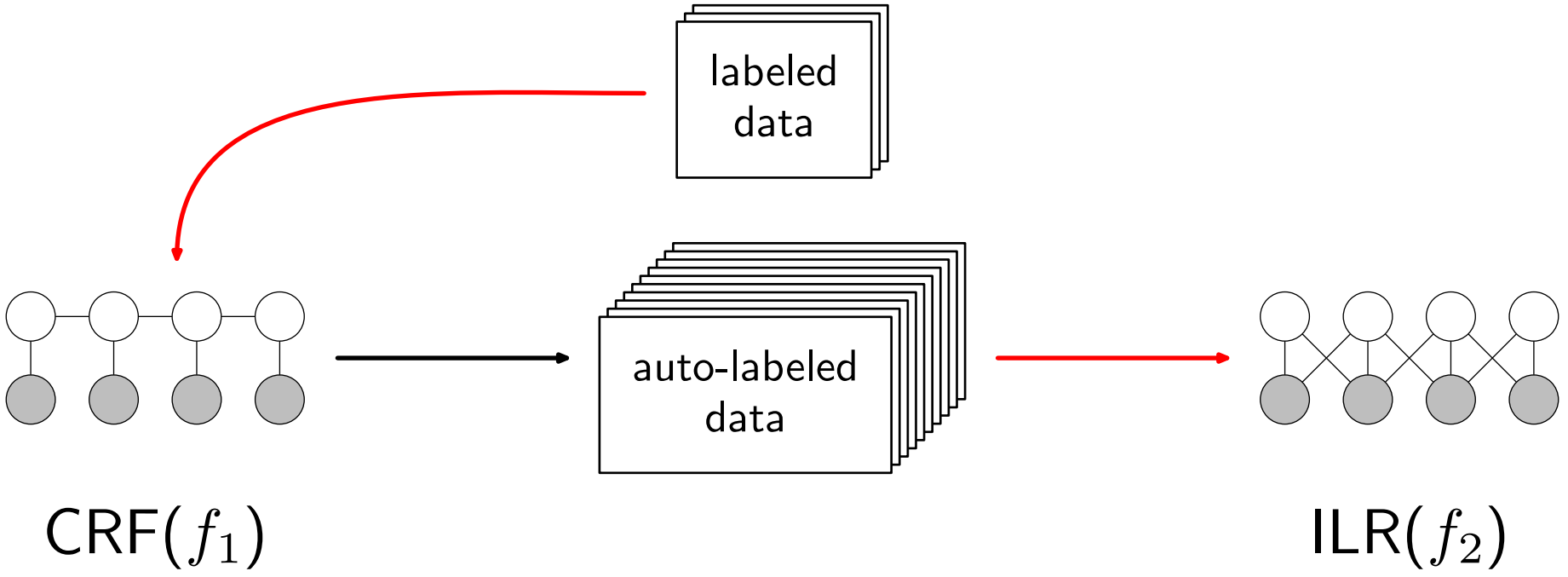
f_2: f_1 applied to larger radius

Structure compilation: reduces the gap between the ILR and CRF

Analysis of structure compilation

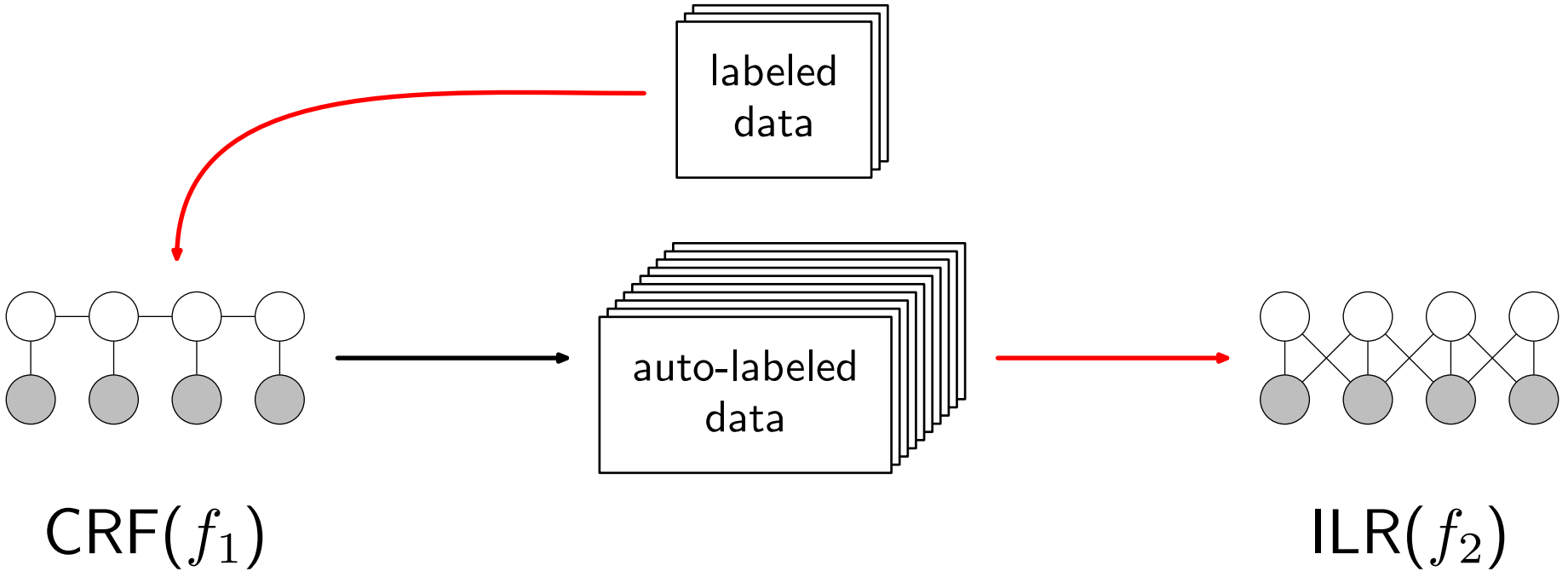


Analysis of structure compilation



Goal: analyze risk of final compiled ILR(f_2)

Analysis of structure compilation

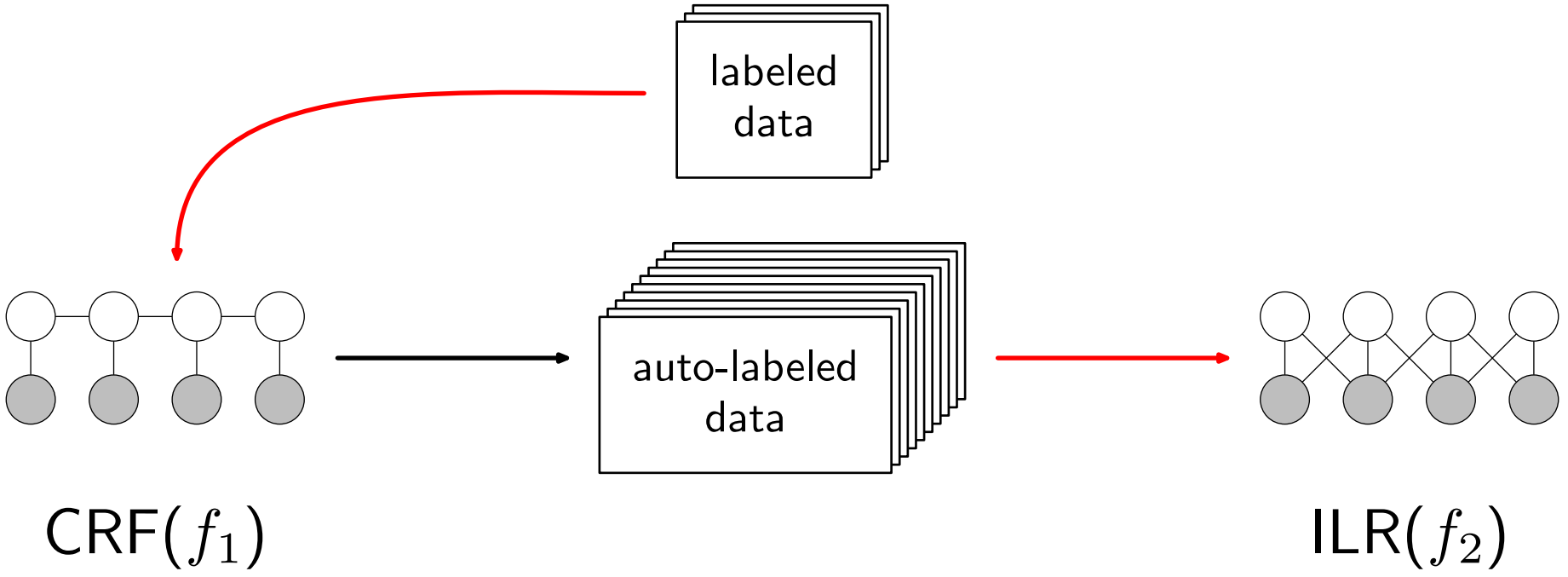


Goal: analyze risk of final compiled ILR(f_2)

Decomposition of errors:

Approximation error: best loss of model (with infinite data)

Analysis of structure compilation

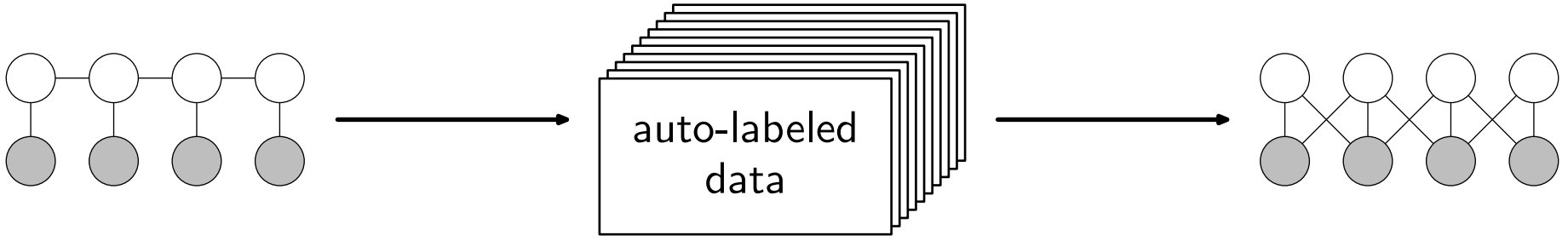


Goal: analyze risk of final compiled ILR(f_2)

Decomposition of errors:

- Approximation error: best loss of model (with infinite data)
- Estimation error: suboptimality due to finite data

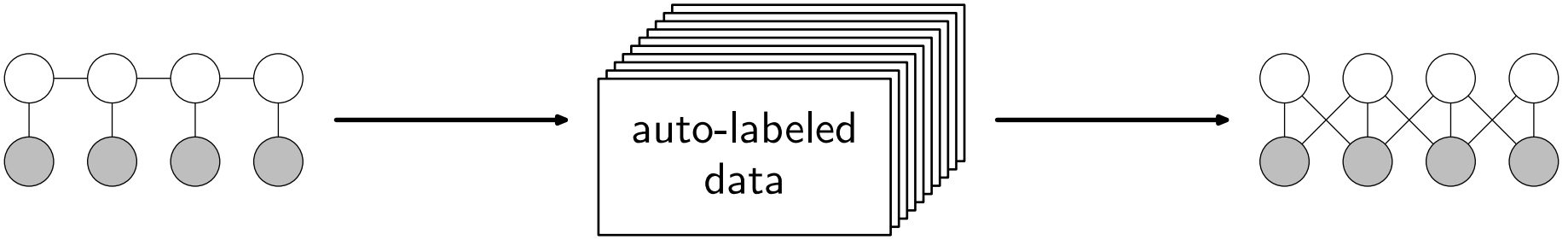
Approximation/estimation errors for ILR



CRF(f_1)

ILR(f_2)

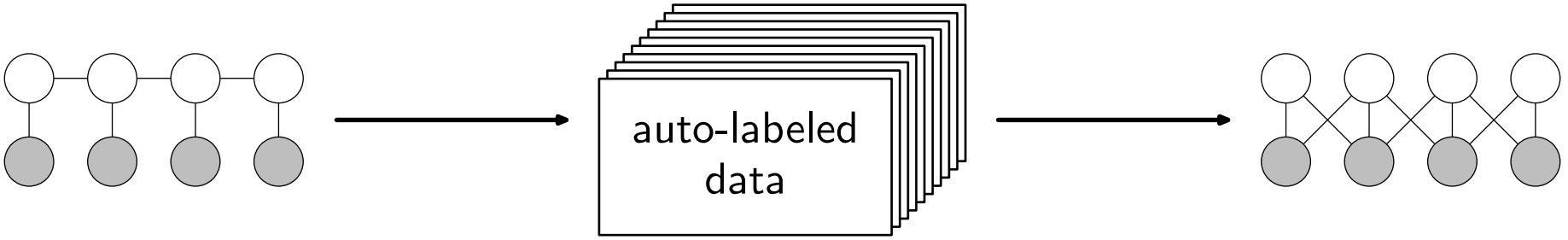
Approximation/estimation errors for ILR



CRF(f_1) ILR(f_2)

p_{C}: CRF(f_1) trained on labeled data

Approximation/estimation errors for ILR

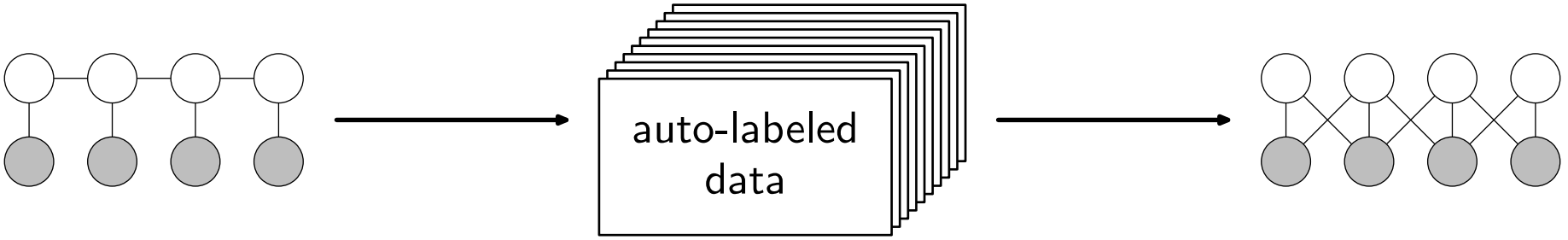


CRF(f_1) ILR(f_2)

p_{C}: CRF(f_1) trained on labeled data

p_{I}: ILR(f_2) trained on m auto-labeled examples

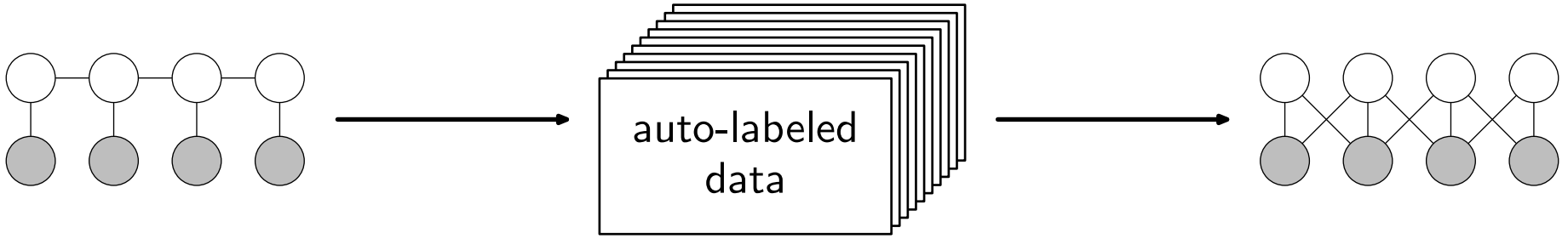
Approximation/estimation errors for ILR



CRF(f_1) ILR(f_2)

p_{C}: CRF(f_1) trained on labeled data
p_{I}: ILR(f_2) trained on m auto-labeled examples
$p_{\mathrm{I}*}$: ILR(f_2) trained on infinite auto-labeled data

Approximation/estimation errors for ILR

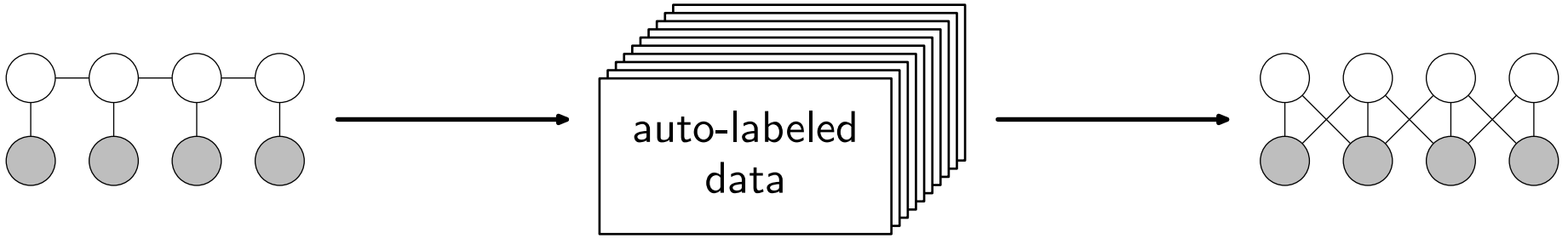


CRF(f_1) ILR(f_2)

p_{C}: CRF(f_1) trained on labeled data
p_{I}: ILR(f_2) trained on m auto-labeled examples
p_{I^*}: ILR(f_2) trained on infinite auto-labeled data

$$\mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{I}}\right) = \underbrace{\mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{I}^*}\right)}_{\text{approx. error}} + \underbrace{\left(\mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{I}}\right) - \mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{I}^*}\right)\right)}_{\text{estimation error}}$$

Approximation/estimation errors for ILR



CRF(f_1) ILR(f_2)

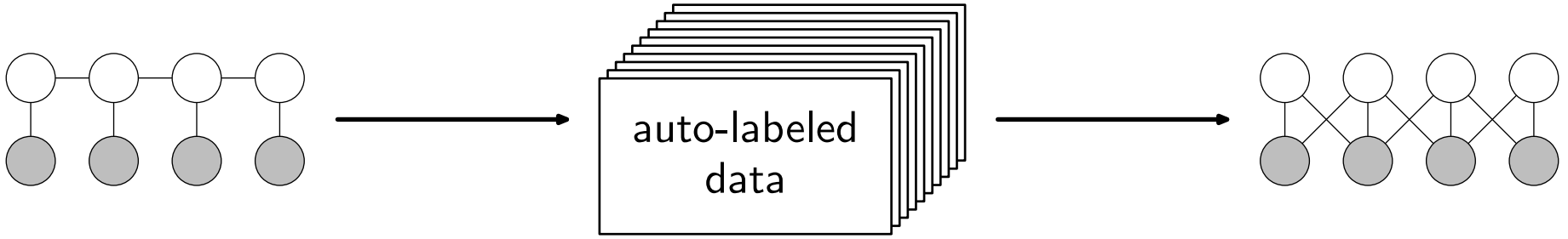
p_{C}: CRF(f_1) trained on labeled data
p_{I}: ILR(f_2) trained on m auto-labeled examples
p_{I^*}: ILR(f_2) trained on infinite auto-labeled data

$$\mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{I}}\right) = \underbrace{\mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{I}^*}\right)}_{\text{approx. error}} + \underbrace{\left(\mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{I}}\right) - \mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{I}^*}\right)\right)}_{\text{estimation error}}$$

Estimation error:

Expected value $= \frac{\#\text{ features}}{m} + o\left(\frac{1}{m}\right) \to 0$ [Liang & Jordan, 2008]

Approximation/estimation errors for ILR



CRF(f_1) ILR(f_2)

p_{C}: CRF(f_1) trained on labeled data
p_{I}: ILR(f_2) trained on m auto-labeled examples
p_{I^*}: ILR(f_2) trained on infinite auto-labeled data

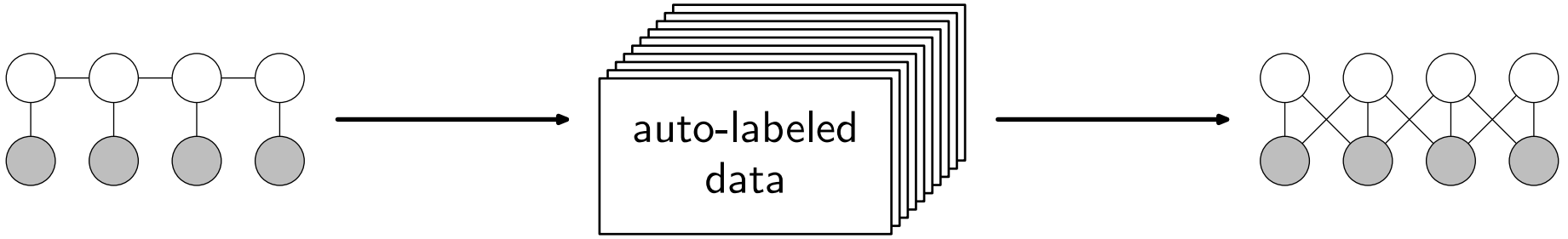
$$\mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{I}}\right) = \underbrace{\mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{I}^*}\right)}_{\text{approx. error}} + \underbrace{\left(\mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{I}}\right) - \mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{I}^*}\right)\right)}_{\text{estimation error}}$$

Estimation error:

Expected value $= \frac{\#\text{ features}}{m} + o\left(\frac{1}{m}\right) \to 0$ [Liang & Jordan, 2008]

Structured compilation can eliminate this error

Approximation/estimation errors for ILR



CRF(f_1) ILR(f_2)

p_{C}: CRF(f_1) trained on labeled data
p_{I}: ILR(f_2) trained on m auto-labeled examples
p_{I^*}: ILR(f_2) trained on infinite auto-labeled data

$$\mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{I}}\right) = \underbrace{\mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{I}^*}\right)}_{\text{approx. error}} + \underbrace{\left(\mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{I}}\right) - \mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{I}^*}\right)\right)}_{\text{estimation error}}$$

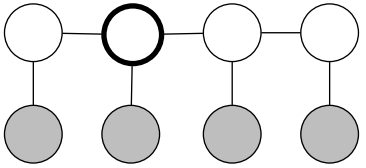
Estimation error:

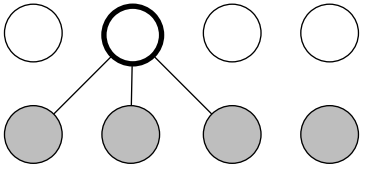
Expected value $= \frac{\#\text{ features}}{m} + o\left(\frac{1}{m}\right) \to 0$ [Liang & Jordan, 2008]

Structured compilation can eliminate this error

Approximation error: next...

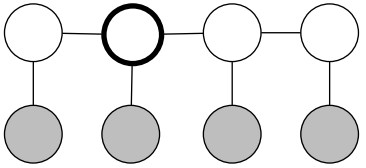
Decomposition of approximation error

CRF(f_1): p_{C} 

ILR(f_2): p_{I^*} 

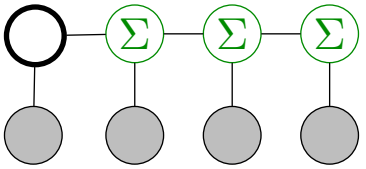
Decomposition of approximation error

CRF(f_1): p_{C}

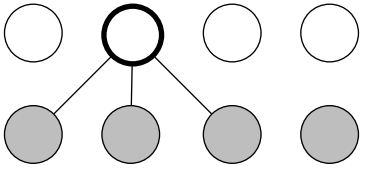


coherence

marginalized CRF(f_1): p_{MC}

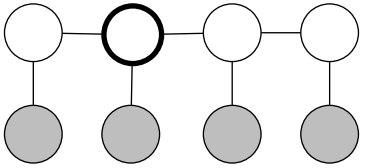


ILR(f_2): $p_{\mathrm{I}*}$



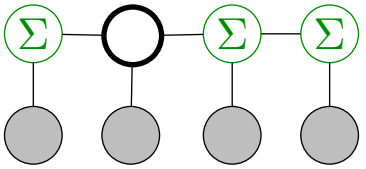
Decomposition of approximation error

CRF(f_1): p_{C}

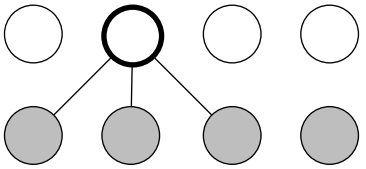


coherence

marginalized CRF(f_1): p_{MC}

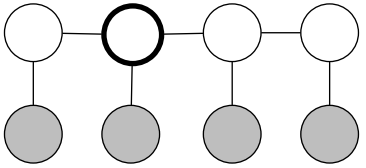


ILR(f_2): p_{I^*}



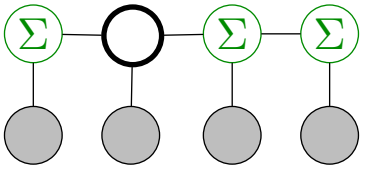
Decomposition of approximation error

CRF(f_1): p_{C}



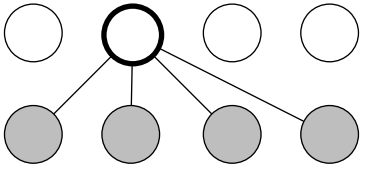
coherence

marginalized CRF(f_1): p_{MC}

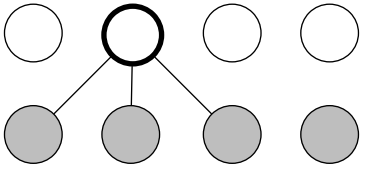


nonlinearities

ILR(f_∞): p_{A^*}

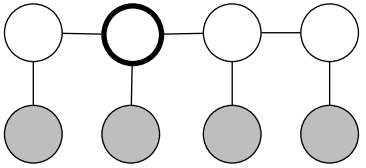


ILR(f_2): p_{I^*}



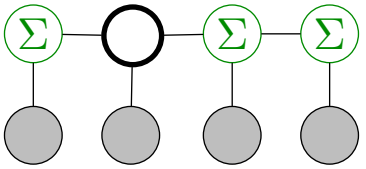
Decomposition of approximation error

CRF(f_1): p_{C}



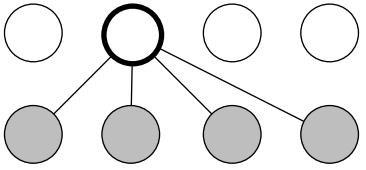
coherence

marginalized CRF(f_1): p_{MC}



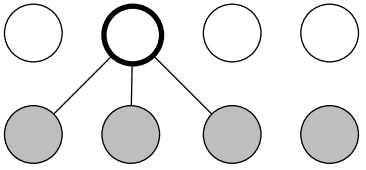
nonlinearities

ILR(f_∞): p_{A^*}



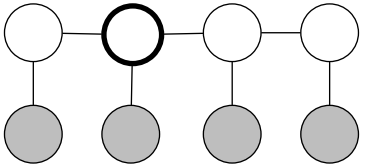
global information

ILR(f_2): p_{I^*}



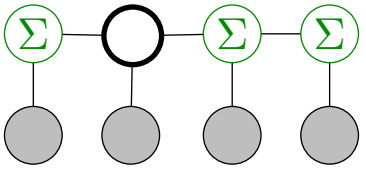
Decomposition of approximation error

CRF(f_1): p_{C}



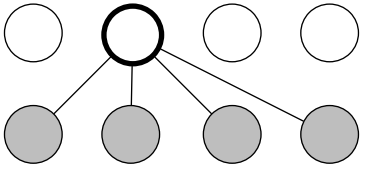
coherence

marginalized CRF(f_1): p_{MC}



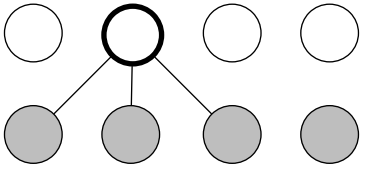
nonlinearities

ILR(f_∞): p_{A^*}



global information

ILR(f_2): p_{I^*}

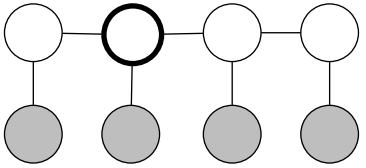


Theorem:

$$\mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{I}^*}\right) = \mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{MC}}\right) + \mathrm{KL}\left(p_{\mathrm{MC}} \,||\, p_{\mathrm{A}^*}\right) + \mathrm{KL}\left(p_{\mathrm{A}^*} \,||\, p_{\mathrm{I}^*}\right)$$

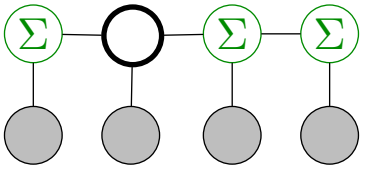
Decomposition of approximation error

CRF(f_1): p_{C}



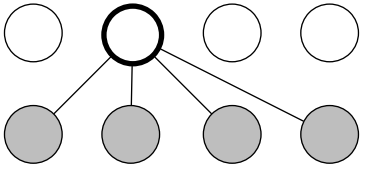
coherence

marginalized CRF(f_1): p_{MC}



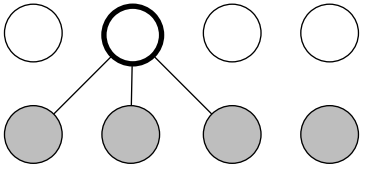
nonlinearities

ILR(f_∞): p_{A^*}



global information

ILR(f_2): p_{I^*}



Theorem:

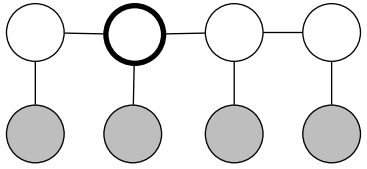
$$\mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{I}^*}\right) = \mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{MC}}\right) + \mathrm{KL}\left(p_{\mathrm{MC}} \,||\, p_{\mathrm{A}^*}\right) + \mathrm{KL}\left(p_{\mathrm{A}^*} \,||\, p_{\mathrm{I}^*}\right)$$

Proof:

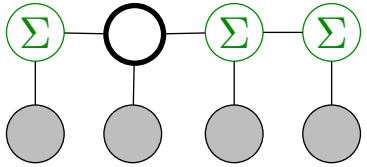
Generalized Pythagorean identity for KL-divergence

Approximation error: coherence

CRF: p_{C}



marginalized CRF: p_{MC}

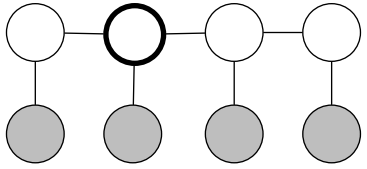


Coherence $= \mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{MC}}\right)$:

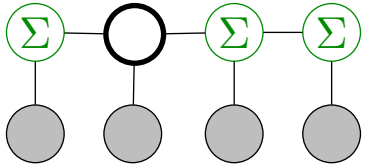
importance of making joint predictions

Approximation error: coherence

CRF: p_{C}



marginalized CRF: p_{MC}



Coherence $= \mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{MC}}\right)$:

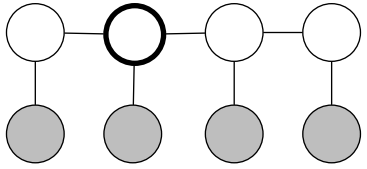
importance of making joint predictions

For a chain CRF:

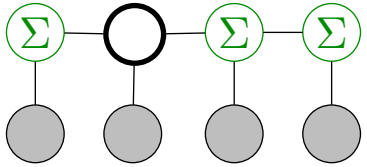
coherence = sum of mutual information along the edges

Approximation error: coherence

CRF: p_{C}



marginalized CRF: p_{MC}



Coherence $= \mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{MC}}\right)$:

importance of making joint predictions

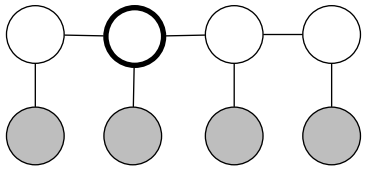
For a chain CRF:

coherence = sum of mutual information along the edges

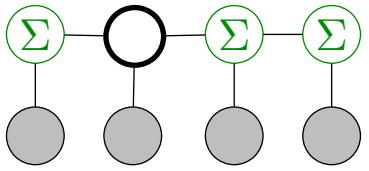
	POS	**NER**
Coherence	0.003	0.009

Approximation error: coherence

CRF: p_{C}



marginalized CRF: p_{MC}



Coherence $= \mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{MC}}\right)$:

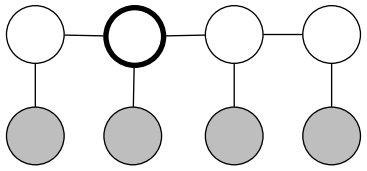
importance of making joint predictions

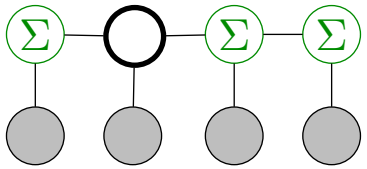
For a chain CRF:

coherence = sum of mutual information along the edges

	POS	**NER**
Coherence	0.003	0.009
Change in accuracy	95.0% $\Rightarrow$ 95.0%	76.3% $\Rightarrow$ 76.0%

Approximation error: coherence

CRF: p_{C} 

marginalized CRF: p_{MC} 

Coherence $= \mathrm{KL}\left(p_{\mathrm{C}} \,||\, p_{\mathrm{MC}}\right)$:

importance of making joint predictions

For a chain CRF:

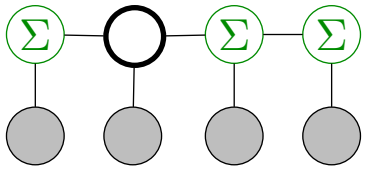
coherence = sum of mutual information along the edges

	POS	**NER**
Coherence	0.003	0.009
Change in accuracy	95.0% ⇒ 95.0%	76.3% ⇒ 76.0%

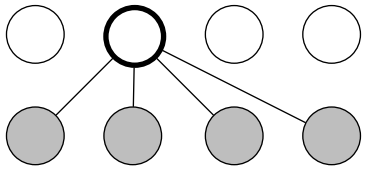
Coherence is not a huge concern (for these applications)

Approximation error: nonlinearities

marginalized CRF: p_{MC}



ILR(f_∞): p_{A^*}

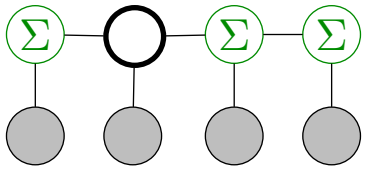


Nonlinearities $= \mathrm{KL}\left(p_{\mathrm{MC}} \,||\, p_{\mathrm{A}^*}\right)$:

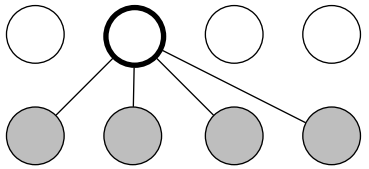
importance of combining features in a nonlinear way

Approximation error: nonlinearities

marginalized CRF: p_{MC}



ILR(f_∞): p_{A^*}



Nonlinearities $= \mathrm{KL}\left(p_{\mathrm{MC}} \,||\, p_{\mathrm{A}^*}\right)$:

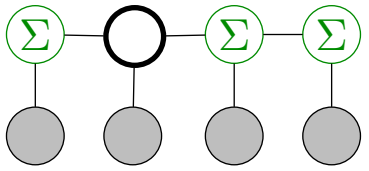
importance of combining features in a nonlinear way

NER experiment:

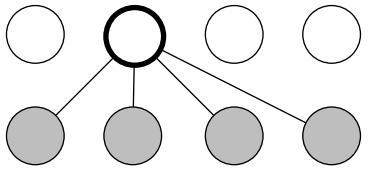
Train a truncated CRF, so that both the truncated CRF (nonlinear) and the ILR (linear) use the **same features**

Approximation error: nonlinearities

marginalized CRF: p_{MC}



ILR(f_∞): p_{A^*}



Nonlinearities $= \mathrm{KL}\left(p_{\mathrm{MC}} \,||\, p_{\mathrm{A}^*}\right)$:

importance of combining features in a nonlinear way

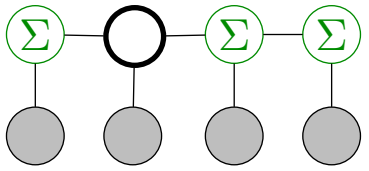
NER experiment:

Train a truncated CRF, so that both the truncated CRF (nonlinear) and the ILR (linear) use the **same features**

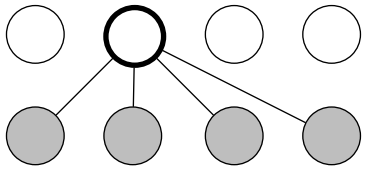
	Truncated CRF	**ILR(f_2)**
Accuracy	76.0%	72.7%

Approximation error: nonlinearities

marginalized CRF: p_{MC}



ILR(f_∞): p_{A^*}



Nonlinearities $= \mathrm{KL}\left(p_{\mathrm{MC}} \,||\, p_{\mathrm{A}^*}\right)$:

importance of combining features in a nonlinear way

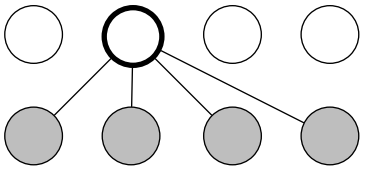
NER experiment:

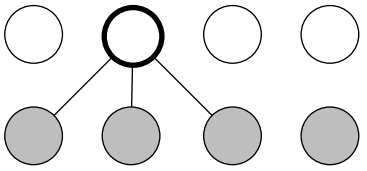
Train a truncated CRF, so that both the truncated CRF (nonlinear) and the ILR (linear) use the **same features**

	Truncated CRF	**ILR(f_2)**
Accuracy	76.0%	72.7%

Nonlinearities play an important role

Approximation error: global information

ILR(f_∞): p_{A^*} 

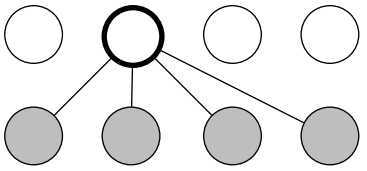
ILR(f_2): p_{I^*} 

Global information $= \mathrm{KL}\left(p_{\mathrm{A}^*} \,||\, p_{\mathrm{I}^*}\right)$:

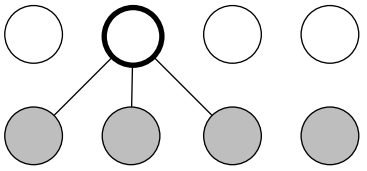
importance of using features on distant parts of the input

Approximation error: global information

ILR(f_∞): p_{A^*}



ILR(f_2): p_{I^*}



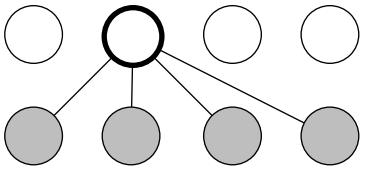
Global information $= \mathrm{KL}\left(p_{\mathrm{A}^*} \,||\, p_{\mathrm{I}^*}\right)$:

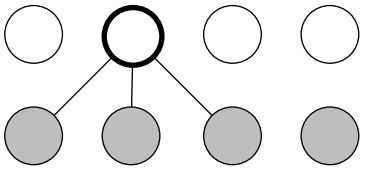
importance of using features on distant parts of the input

NER experiment:

Compare truncated CRF with marginalized CRF (they differ only in the features used)

Approximation error: global information

ILR(f_∞): p_{A^*} 

ILR(f_2): p_{I^*} 

Global information $= \mathrm{KL}\left(p_{\mathrm{A}^*} \,||\, p_{\mathrm{I}^*}\right)$:

importance of using features on distant parts of the input

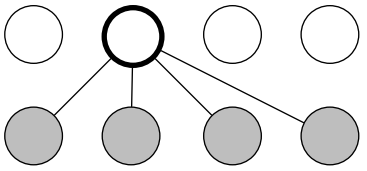
NER experiment:

Compare truncated CRF with marginalized CRF (they differ only in the features used)

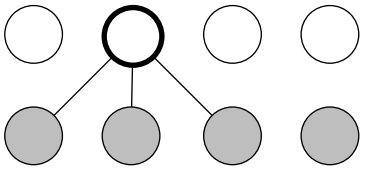
	Marginalized CRF	**Truncated CRF**
Accuracy	76.0%	76.0%

Approximation error: global information

ILR(f_∞): p_{A^*}



ILR(f_2): p_{I^*}



Global information $= \mathrm{KL}\left(p_{\mathrm{A}^*} \,||\, p_{\mathrm{I}^*}\right)$:

importance of using features on distant parts of the input

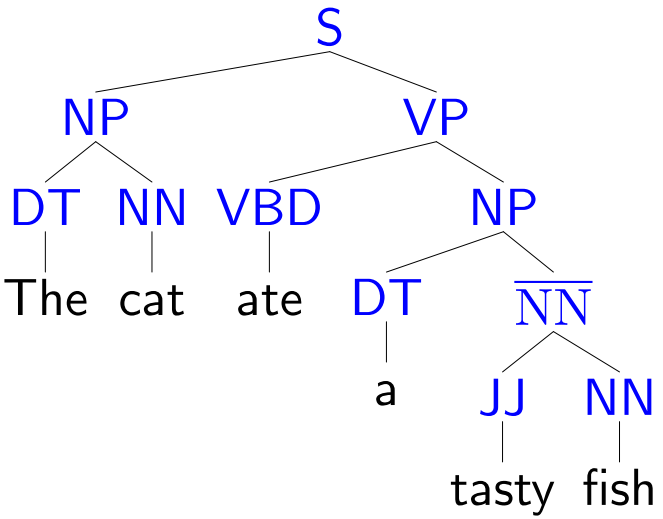
NER experiment:

Compare truncated CRF with marginalized CRF (they differ only in the features used)

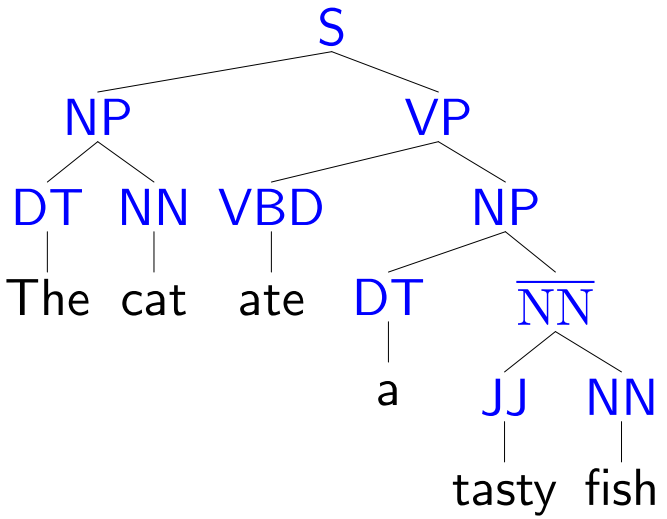
	Marginalized CRF	**Truncated CRF**
Accuracy	76.0%	76.0%

Distant information is not essential (for these applications)

Structure compilation for parsing

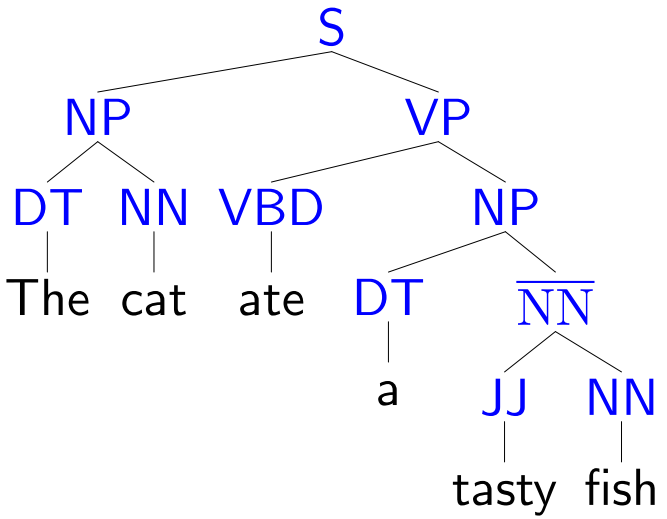


Structure compilation for parsing



Sentence length: ℓ
Number of grammar symbols: K
Number of grammar rules: $G \gg \ell, K$

Structure compilation for parsing



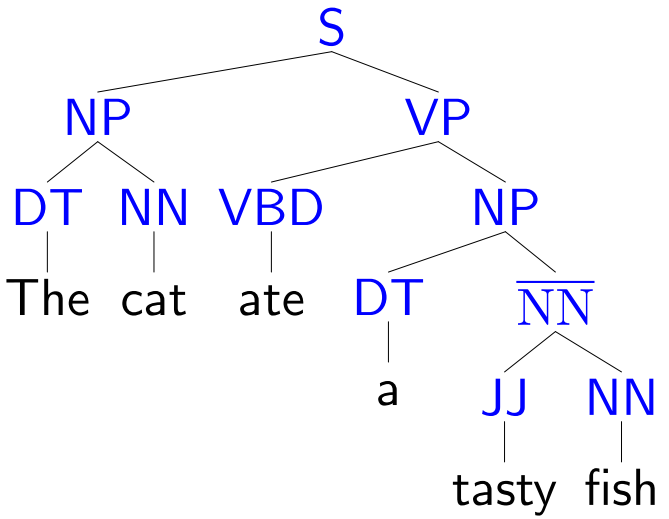
Sentence length: ℓ
Number of grammar symbols: K
Number of grammar rules: $G \gg \ell, K$

Parse time/sentence

Structured model: $O(\ell^3 G)$

Standard dynamic program for context-free grammars

Structure compilation for parsing



Sentence length: ℓ
Number of grammar symbols: K
Number of grammar rules: $G \gg \ell, K$

Parse time/sentence

Structured model: $O(\ell^3 G)$

Standard dynamic program for context-free grammars

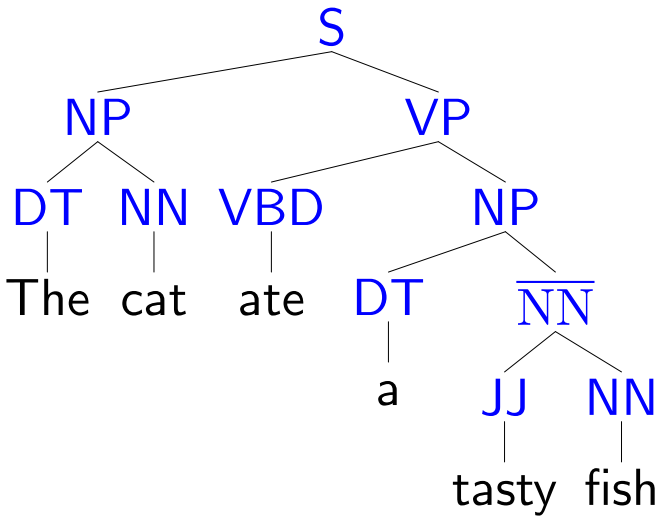
Independent model: $O(\ell^3 + K\ell^2)$

For each of $O(\ell^2)$ spans:

Make a soft prediction of whether it's a constituent

(features: words/tags/prefixes/suffixes on entire span)

Structure compilation for parsing



Sentence length: ℓ
Number of grammar symbols: K
Number of grammar rules: $G \gg \ell, K$

Parse time/sentence

Structured model: $O(\ell^3 G)$

Standard dynamic program for context-free grammars

Independent model: $O(\ell^3 + K\ell^2)$

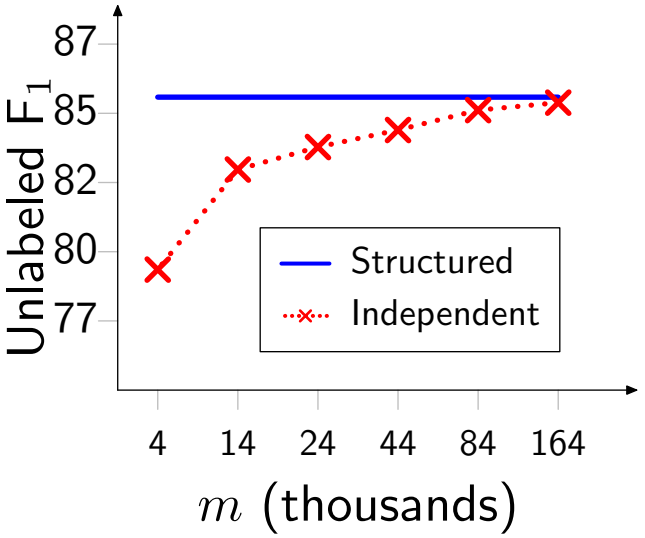
For each of $O(\ell^2)$ spans:

Make a soft prediction of whether it's a constituent

(features: words/tags/prefixes/suffixes on entire span)

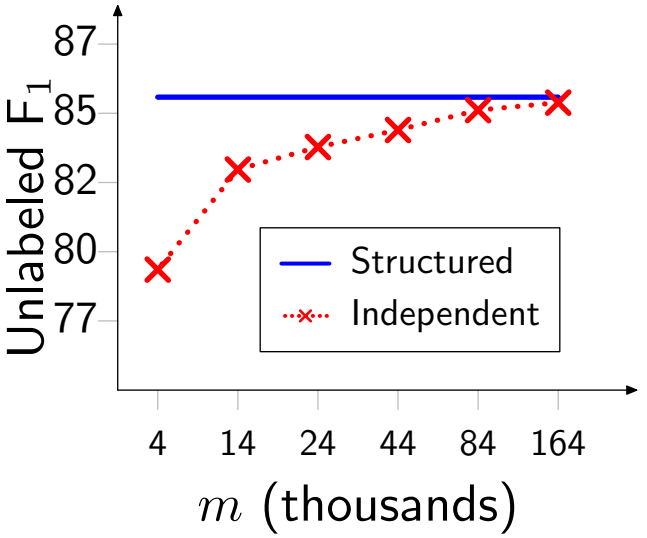
Run a dynamic program to choose the best tree

Parsing results

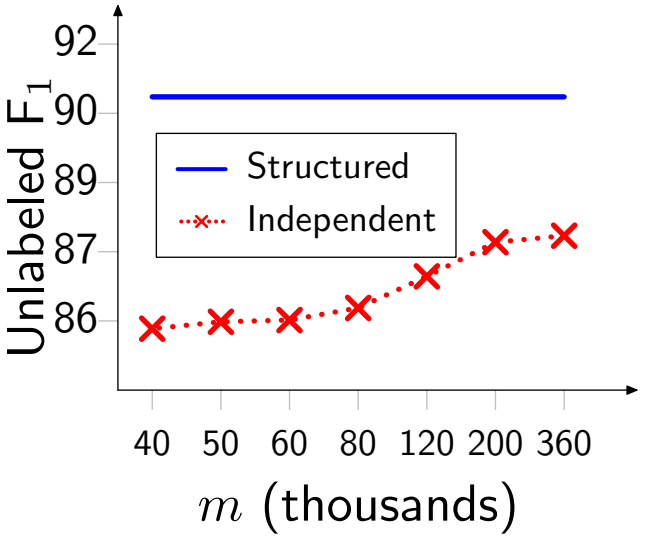


4K labeled sentences

Parsing results

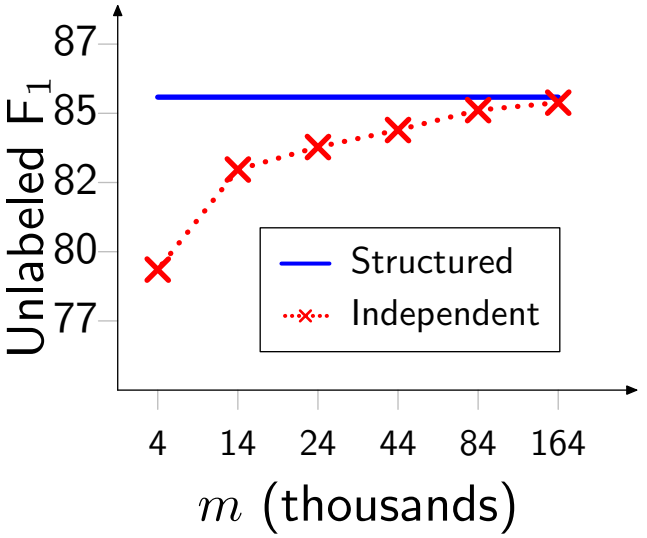


4K labeled sentences

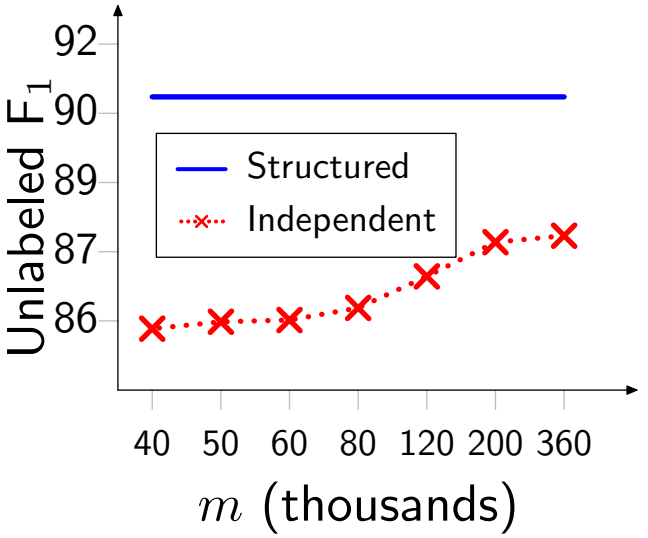


40K labeled sentences

Parsing results



4K labeled sentences



40K labeled sentences

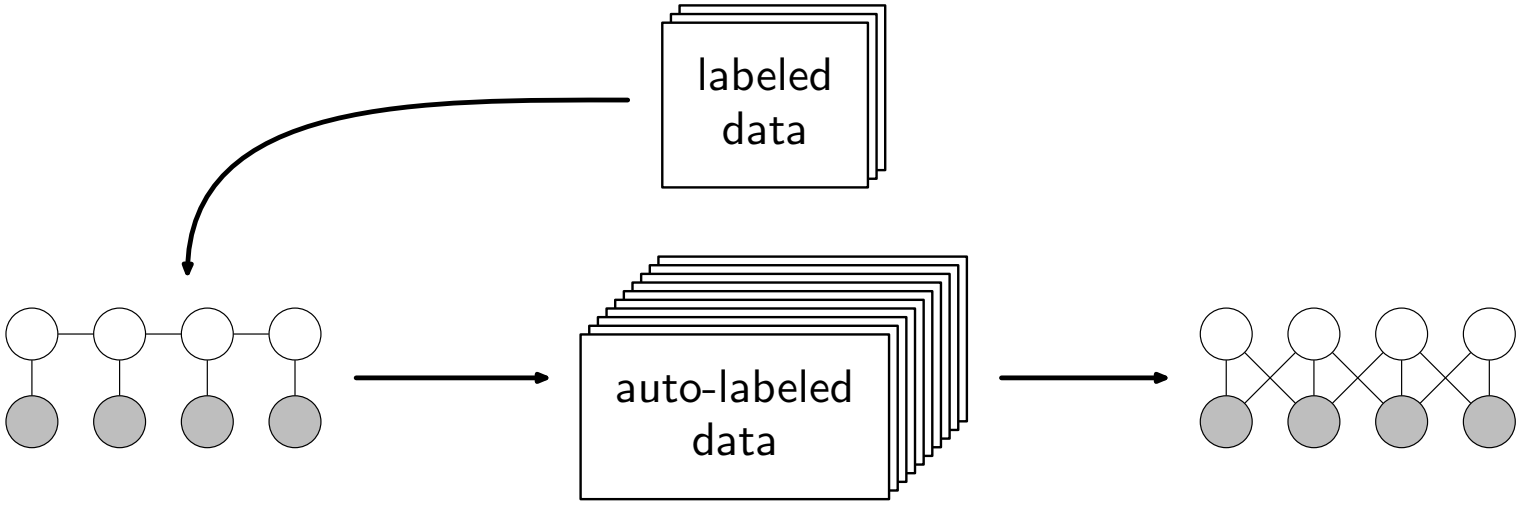
- Structure is important in parsing
- Need richer features or nonlinearities for the independent model to catch up

Summary of structure compilation

Motivation: want fast CRF-level accuracy at test time

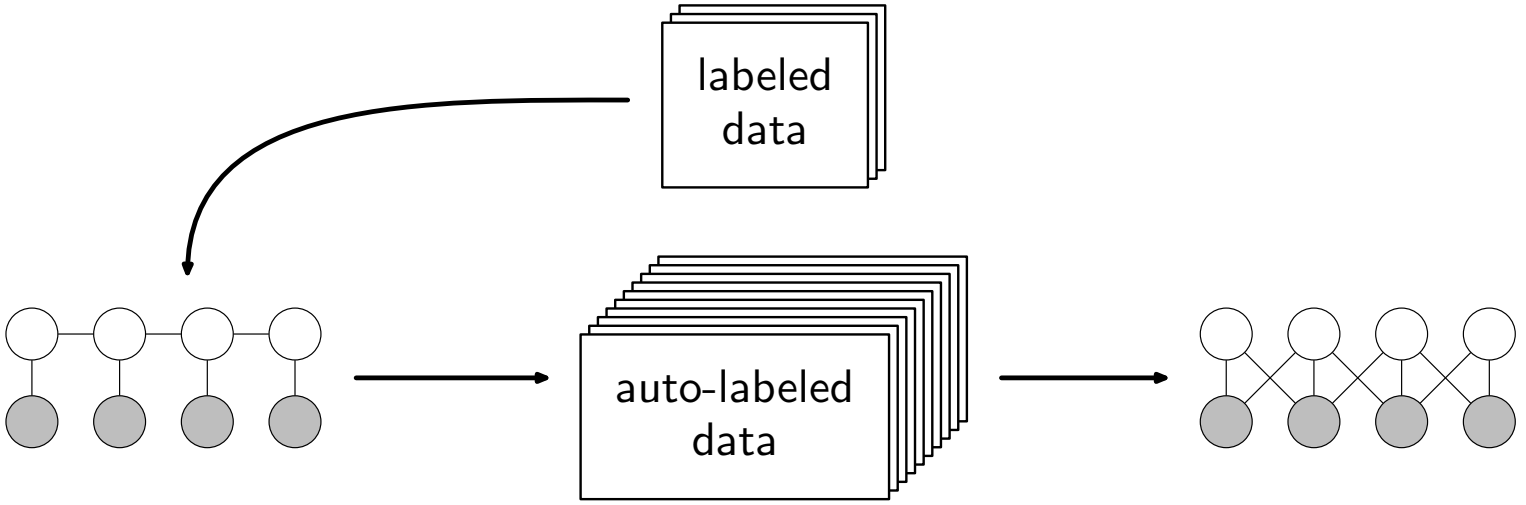
Summary of structure compilation

Motivation: want fast CRF-level accuracy at test time



Summary of structure compilation

Motivation: want fast CRF-level accuracy at test time

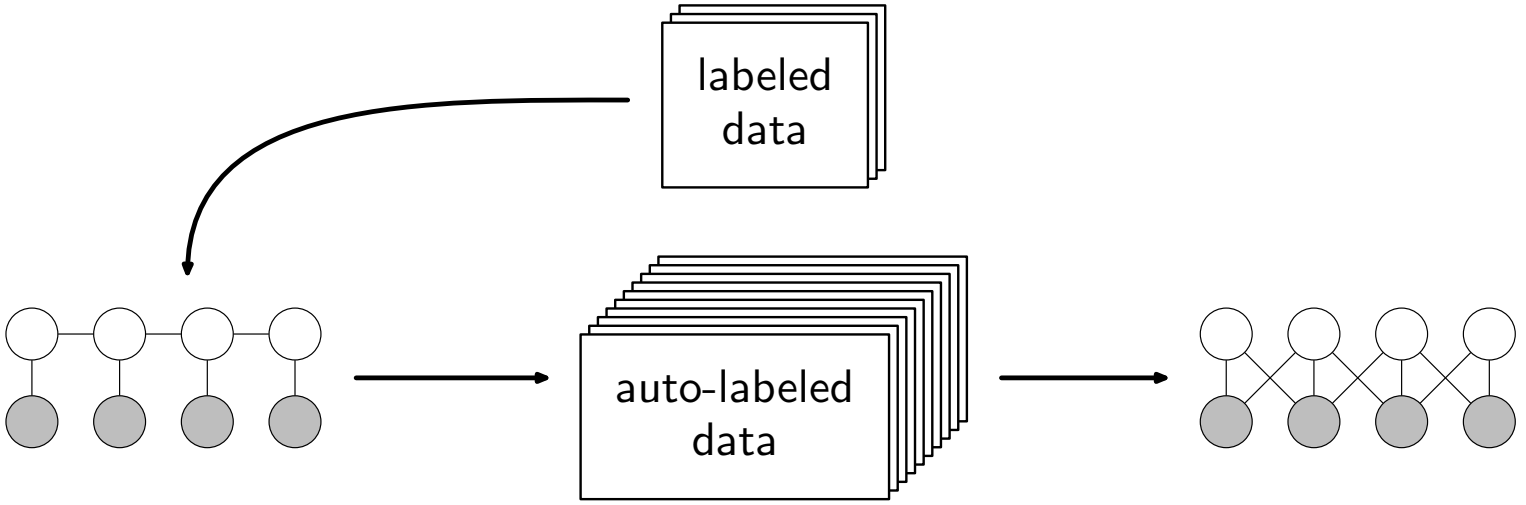


CRF(f_1)
computationally complex

ILR(f_2)
computationally simple

Summary of structure compilation

Motivation: want fast CRF-level accuracy at test time



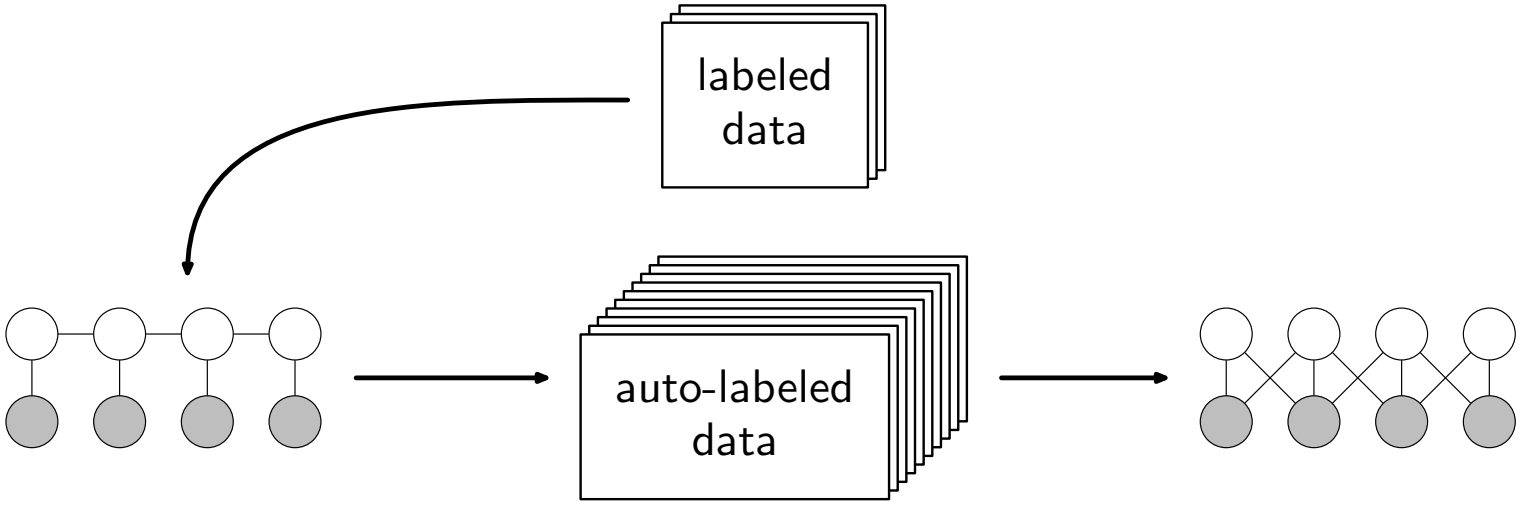
CRF(f_1)
computationally complex
statistically simple

ILR(f_2)
computationally simple
statistically complex

Estimation error: structure compilation can easily drive it to 0

Summary of structure compilation

Motivation: want fast CRF-level accuracy at test time



CRF(f_1)
computationally complex
statistically simple
very expressive

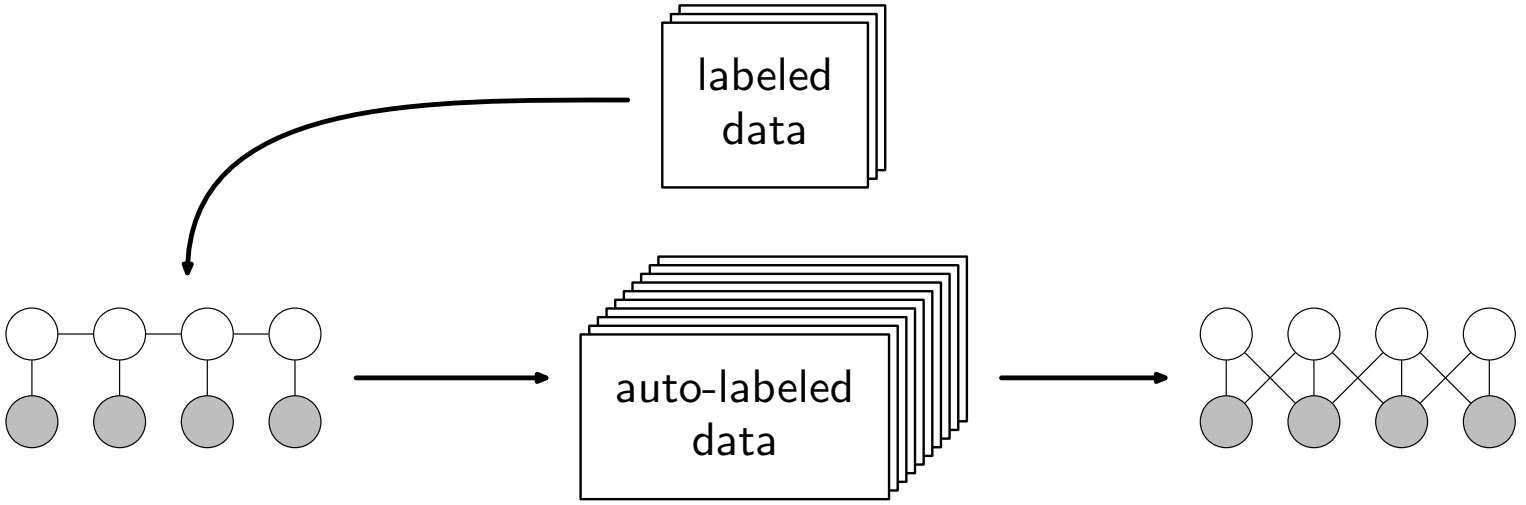
ILR(f_2)
computationally simple
statistically complex
not as expressive

Estimation error: structure compilation can easily drive it to 0

Approximation error: advantages of CRF over ILR

Summary of structure compilation

Motivation: want fast CRF-level accuracy at test time



CRF(f_1)
computationally complex
statistically simple
very expressive

ILR(f_2)
computationally simple
statistically complex
not as expressive

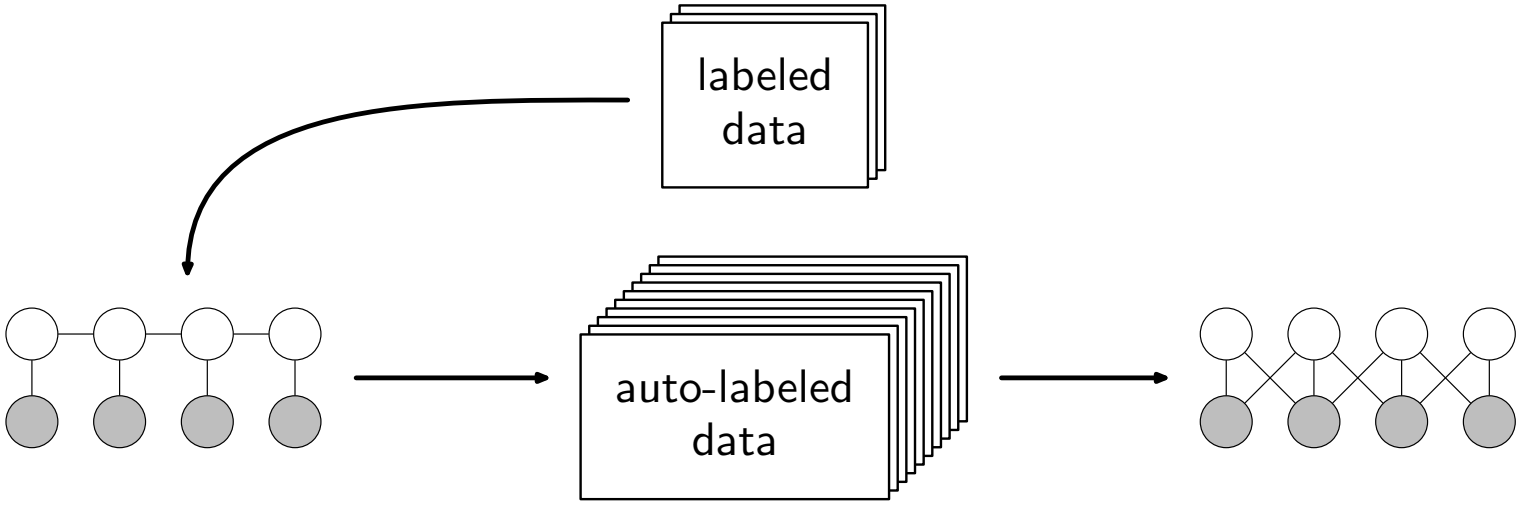
Estimation error: structure compilation can easily drive it to 0

Approximation error: advantages of CRF over ILR

- ILR needs rich features to compensate

Summary of structure compilation

Motivation: want fast CRF-level accuracy at test time



CRF(f_1)
computationally complex
statistically simple
very expressive

ILR(f_2)
computationally simple
statistically complex
not as expressive

Estimation error: structure compilation can easily drive it to 0

Approximation error: advantages of CRF over ILR

- ILR needs rich features to compensate
- CRF's nonlinearities are important