Learning a Part-of-Speech Tagger from Two Hours of Annotation

Dan Garrette and Jason Baldridge

University of Texas at Austin

Low-Resource Languages

6,900 languages in the world

~30 have non-negligible quantities of data

No million-word corpus for any endangered language

[Maxwell and Hughes, 2006]
[Abney and Bird, 2010]

Low-Resource Languages

Kinyarwanda

Niger-Congo; morphologically-rich

Malagasy

Austronesian; spoken in Madagascar

Also, English

Low-Resource Languages

Supervised training is not an option.

We do semi-supervised training.

➡ Annotate some data by hand

... cheaply

... like, in 2 hours

Semi-Supervised Training

HMM with Expectation-Maximization (EM)

Need:

- Large **raw** corpus ← know how to get this
- Tag dictionary ← where is this from?

[Kupiec, 1992]
[Merialdo, 1994]

Tag Dictionary

Most previous work:

Extract from a **large labeled corpus**

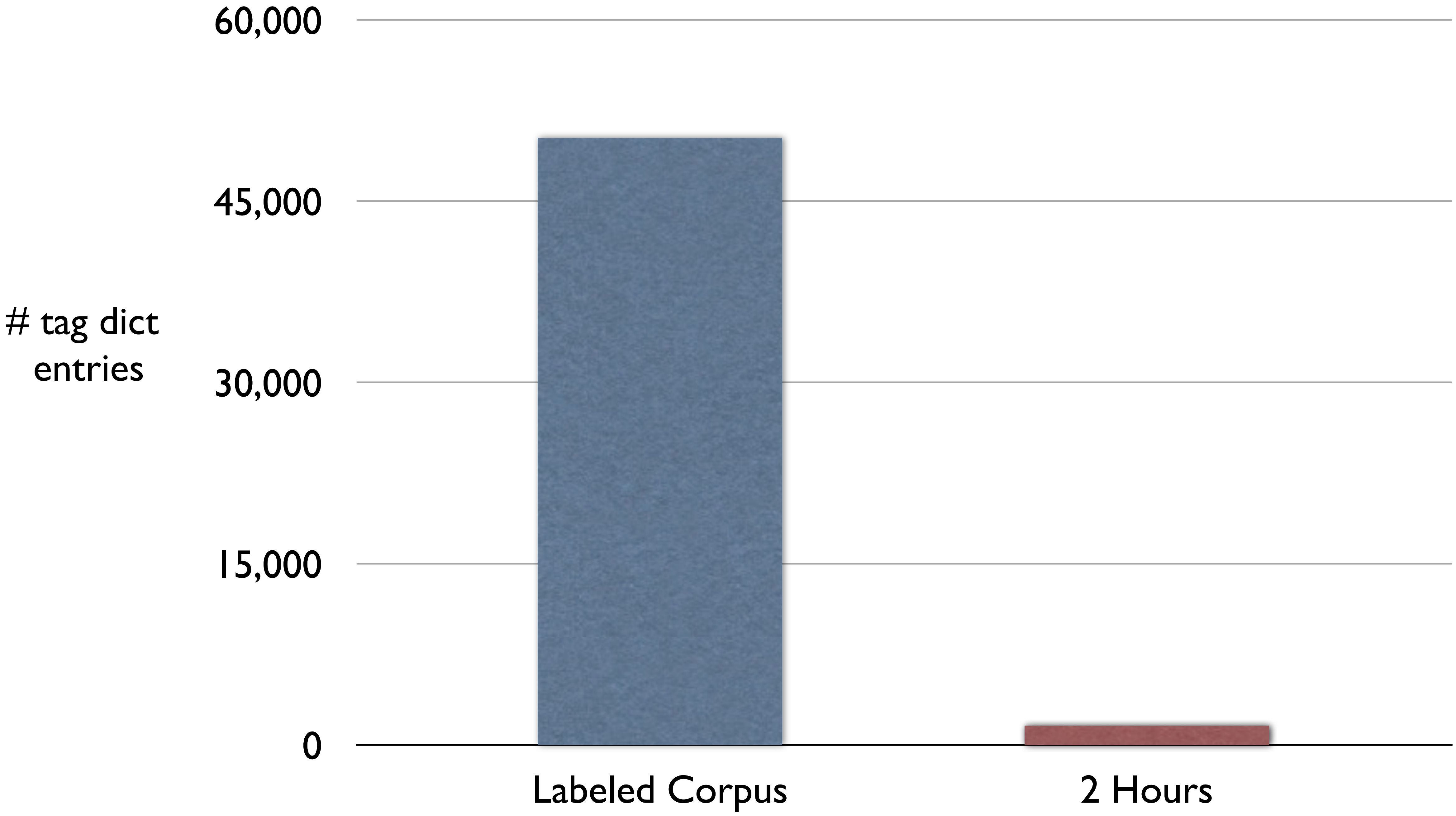
→ too **complete**

→ too **clean**

→ too **biased**

unrealistic

A **Real** Tag Dictionary

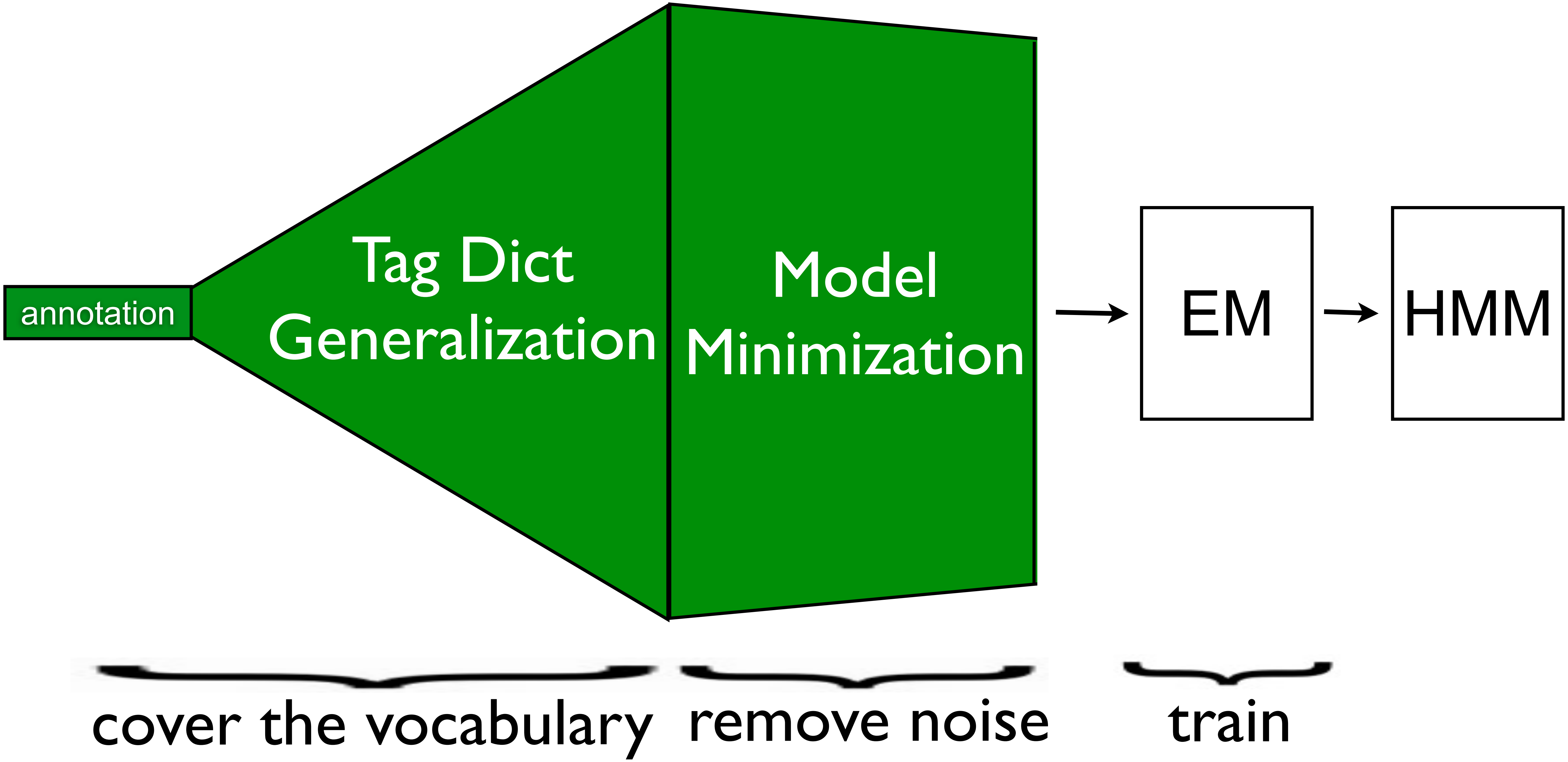


A **Real** Tag Dictionary

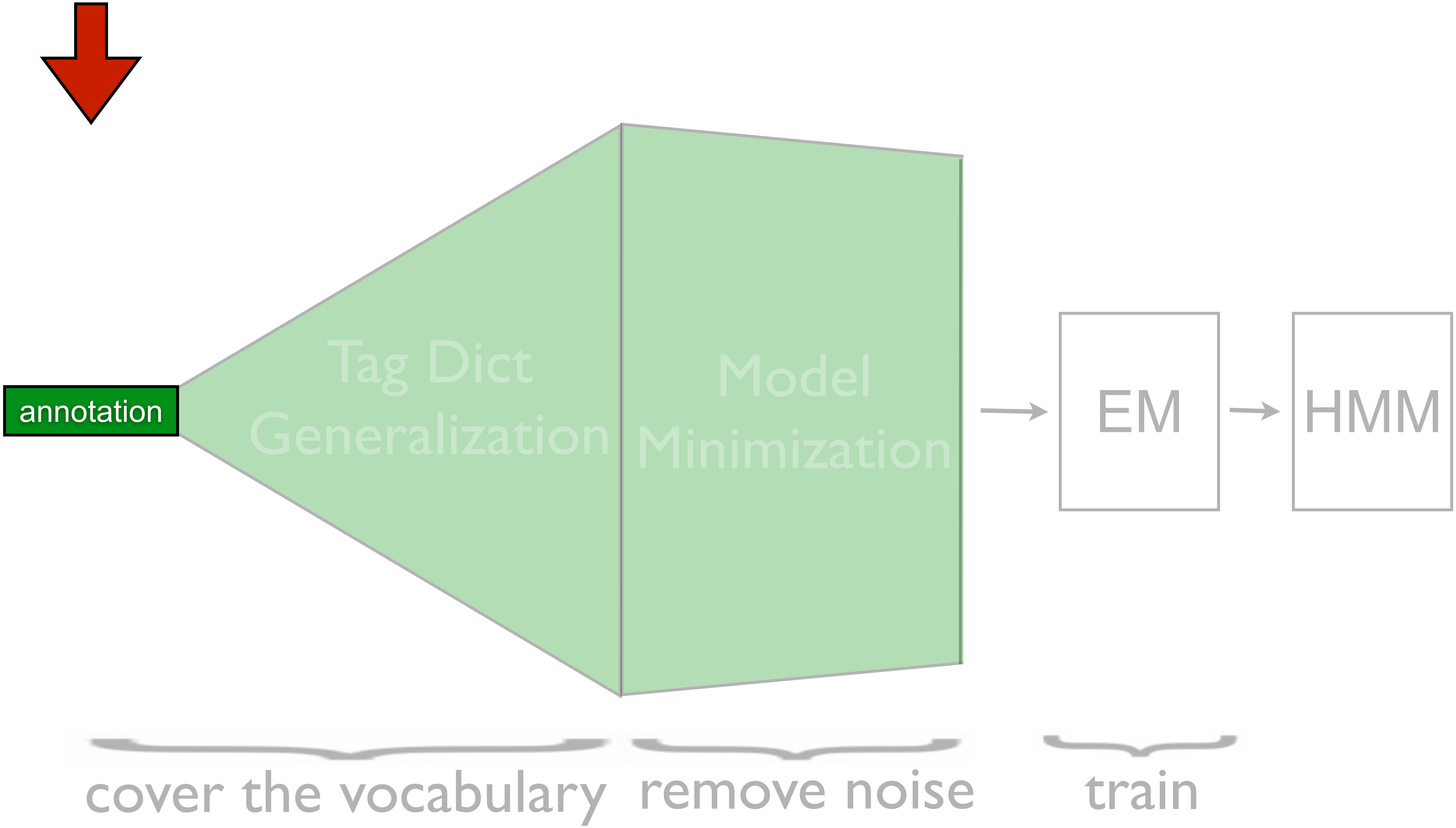
Extremely low coverage means most words are unknown

⇒ **Bad for EM** (poorly constrained)

Our Approach



Our Approach



Collecting Annotations

Task #1 -- **2 hours** to create a **tag dictionary**

ordered by frequency

,	,		
the	DT		
.	.		
of	IN	RP	
to	TO	RP	
a	DT		
and	CC		
⋮	⋮		
only	RB		
can	VB	VBP	MD
York	NNP		
into	IN	RP	
after	IN	RP	
president	NN		
⋮	⋮		

Collecting Annotations

Task #2 -- **2 hours** to annotate **full sentences**

Pierre	Vinken	,	61	years	old	,	will	join	the	board	as	a	nonexecutive	director	Nov.	29	.
NNP	NNP	,	CD	NNS	JJ	,	MD	VB	DT	NN	IN	DT	JJ	NN	NNP	CD	.

Mr.	Vinken	is	chairman	of	Elsevier	N.V.	,	the	Dutch	publishing	group	.
NNP	NNP	VB	NN	IN	NNP	NNP	,	DT	JJ	JJ	NN	.

⋮

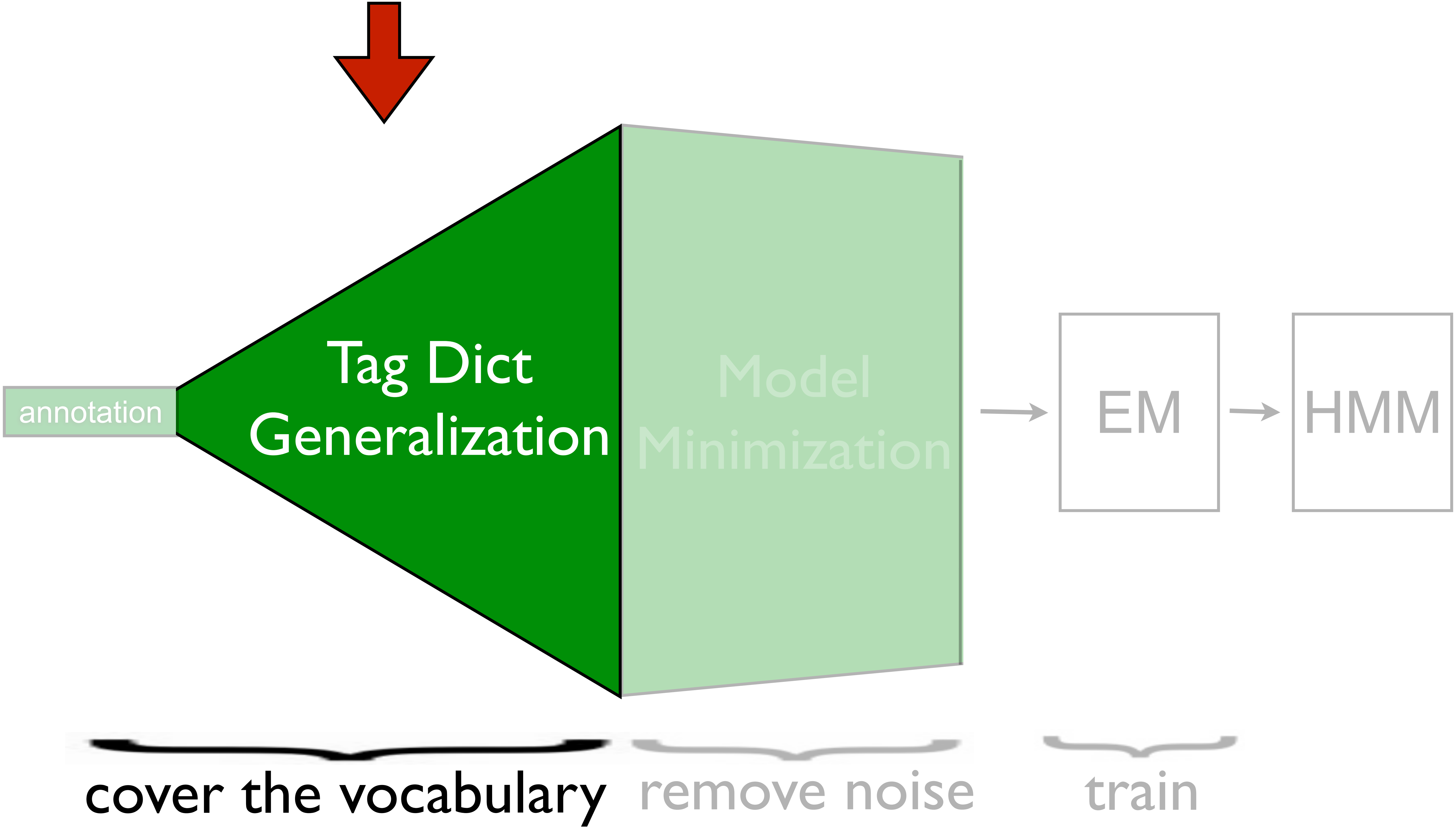
Collecting Annotations

In 2 hours:

	# sent	# tok	# TD entries
Full Sentences	90	1537	**750**
Tag Dict			**1798**

(for Kinyarwanda)

Our Approach



Tag Dict Generalization

These annotations are too sparse!

➡ Generalize to the entire vocabulary

Tag Dict Generalization

Haghighi and Klein (2006) do this with
a vector space.

We don’t have enough raw data

Das and Petrov (2011) do this with
a parallel corpus.

We don’t have a parallel corpus

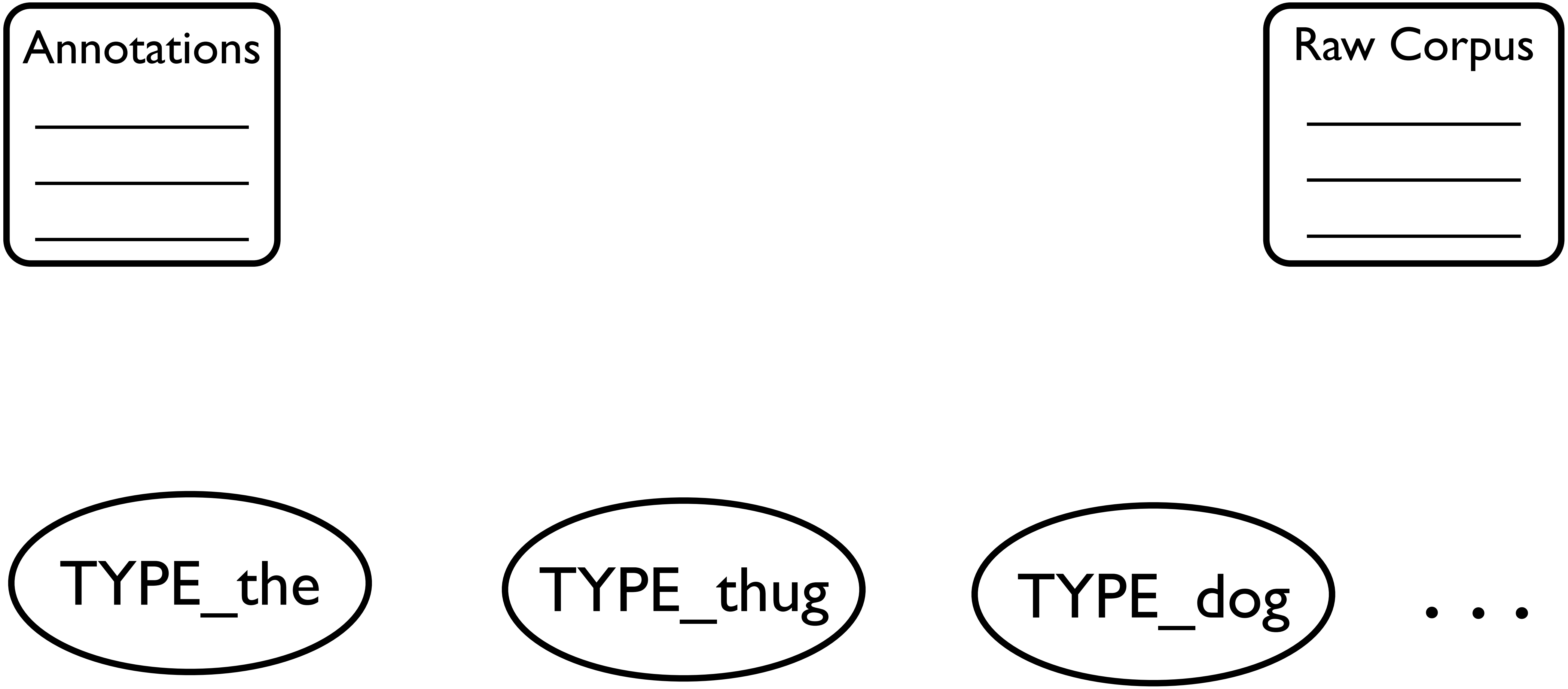
Tag Dict Generalization

Our strategy: Label Propagation

- **Connect** annotations to raw corpus tokens
- Push tag labels to **entire corpus**

[Talukdar and Crammer. 2009]

Tag Dict Generalization



Annotations

Raw Corpus

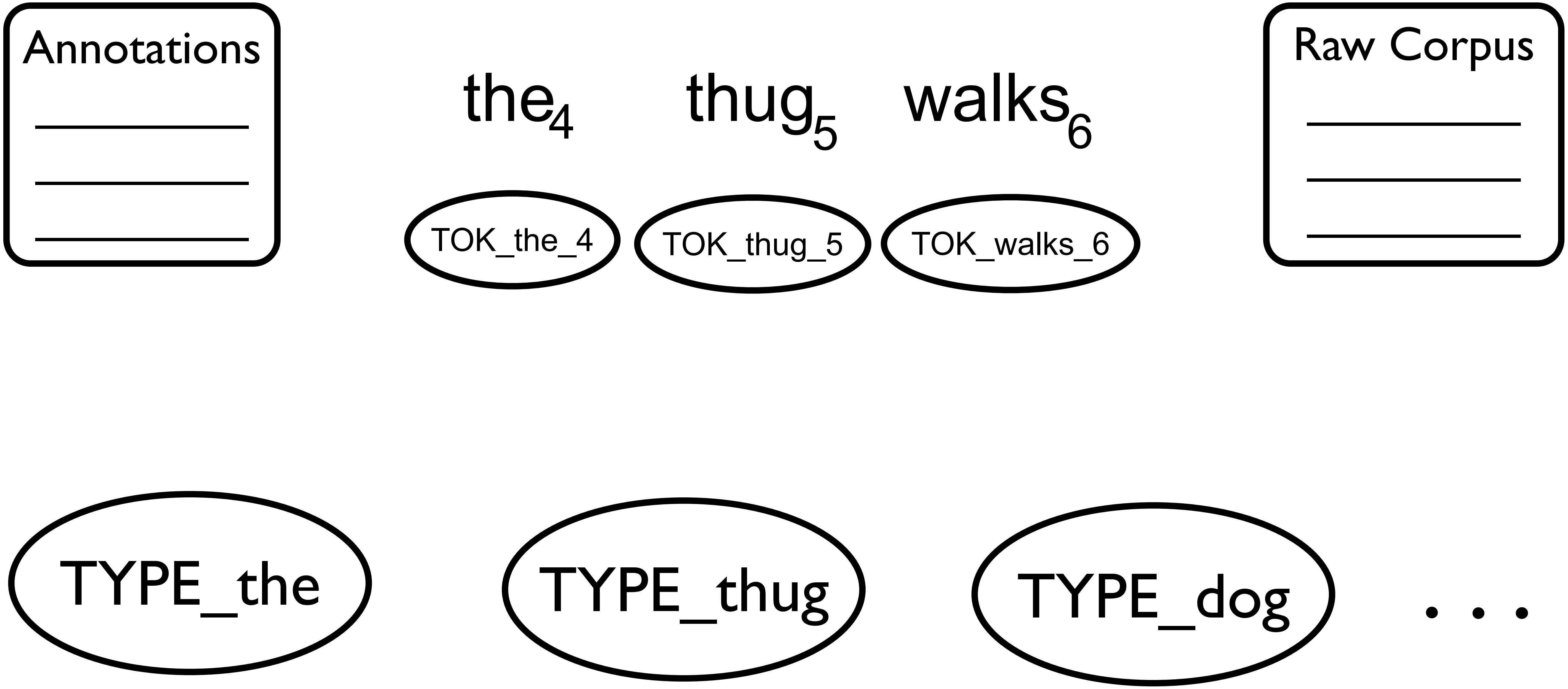
TYPE_the

TYPE_thug

TYPE_dog

. . .

Tag Dict Generalization



Annotations

the$_4$ thug$_5$ walks$_6$

TOK_the_4

TOK_thug_5

TOK_walks_6

Raw Corpus

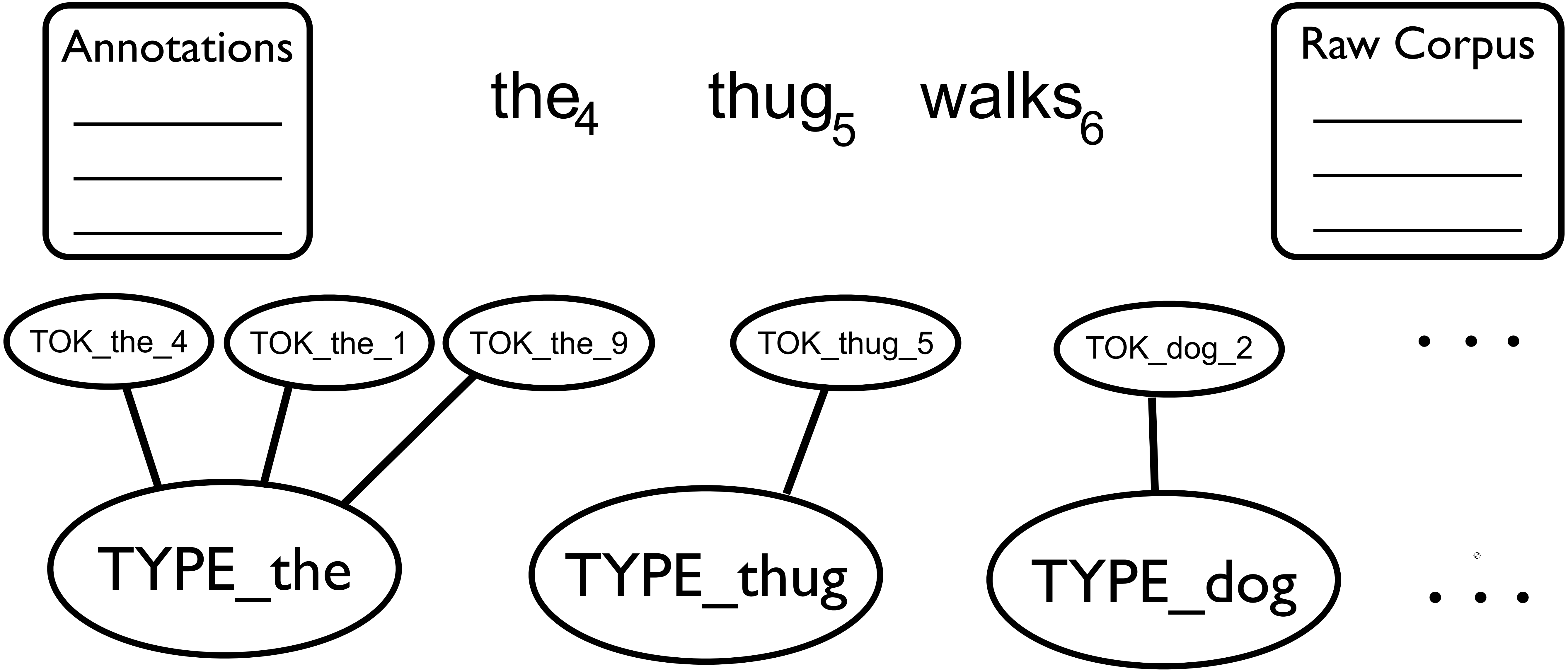
TYPE_the

TYPE_thug

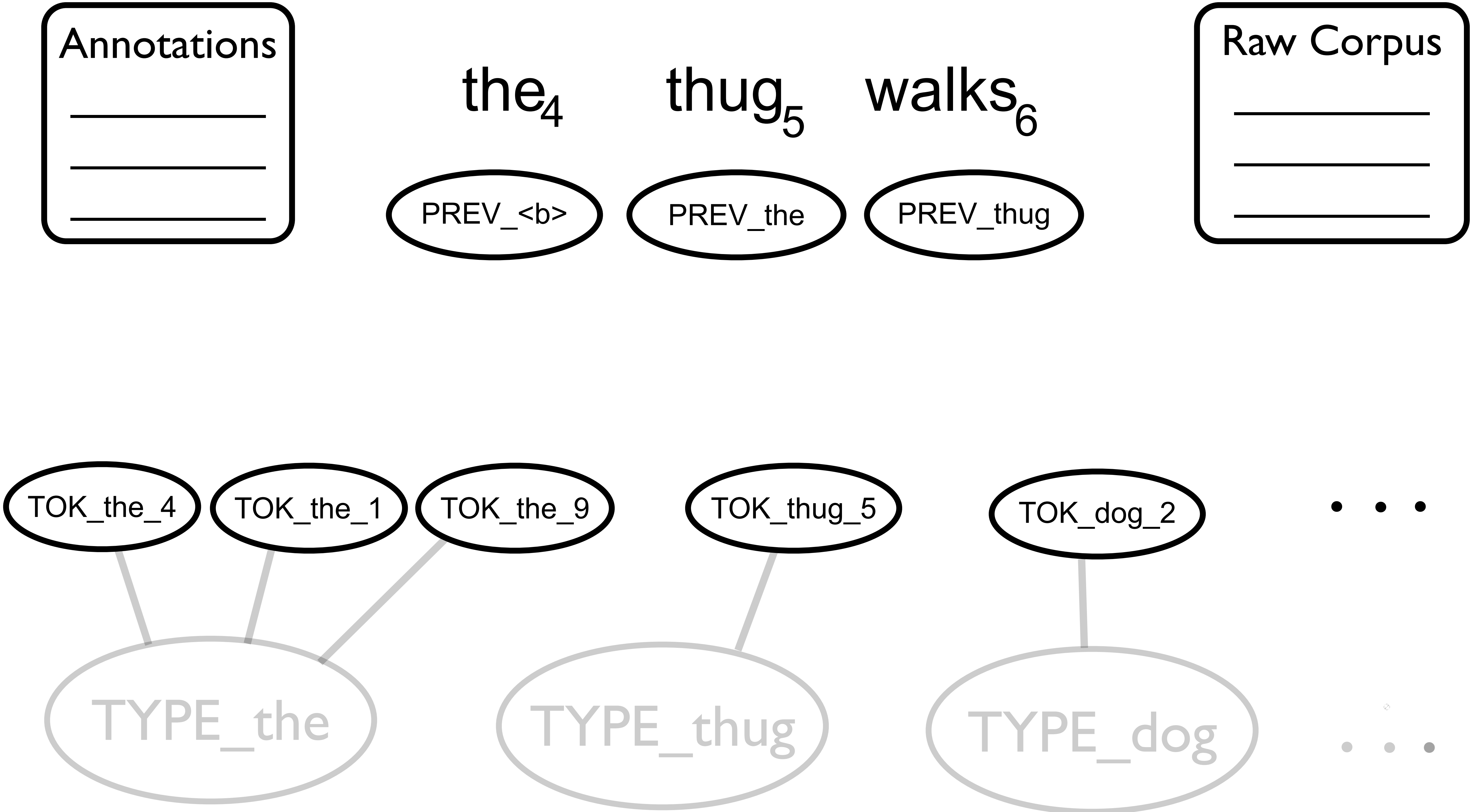
TYPE_dog

. . .

Tag Dict Generalization



Tag Dict Generalization



Annotations

the$_4$ thug$_5$ walks$_6$

PREV_<b>

PREV_the

PREV_thug

Raw Corpus

TOK_the_4

TOK_the_1

TOK_the_9

TOK_thug_5

TOK_dog_2

. . .

TYPE_the

TYPE_thug

TYPE_dog

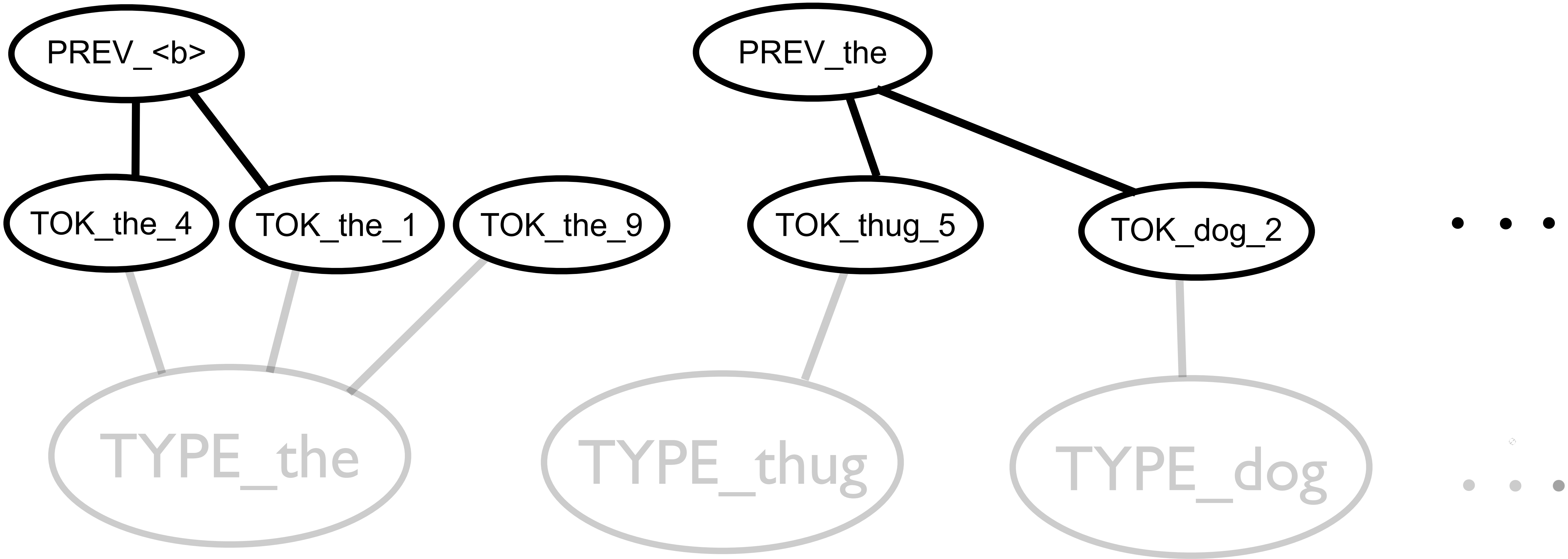
. . .

Tag Dict Generalization

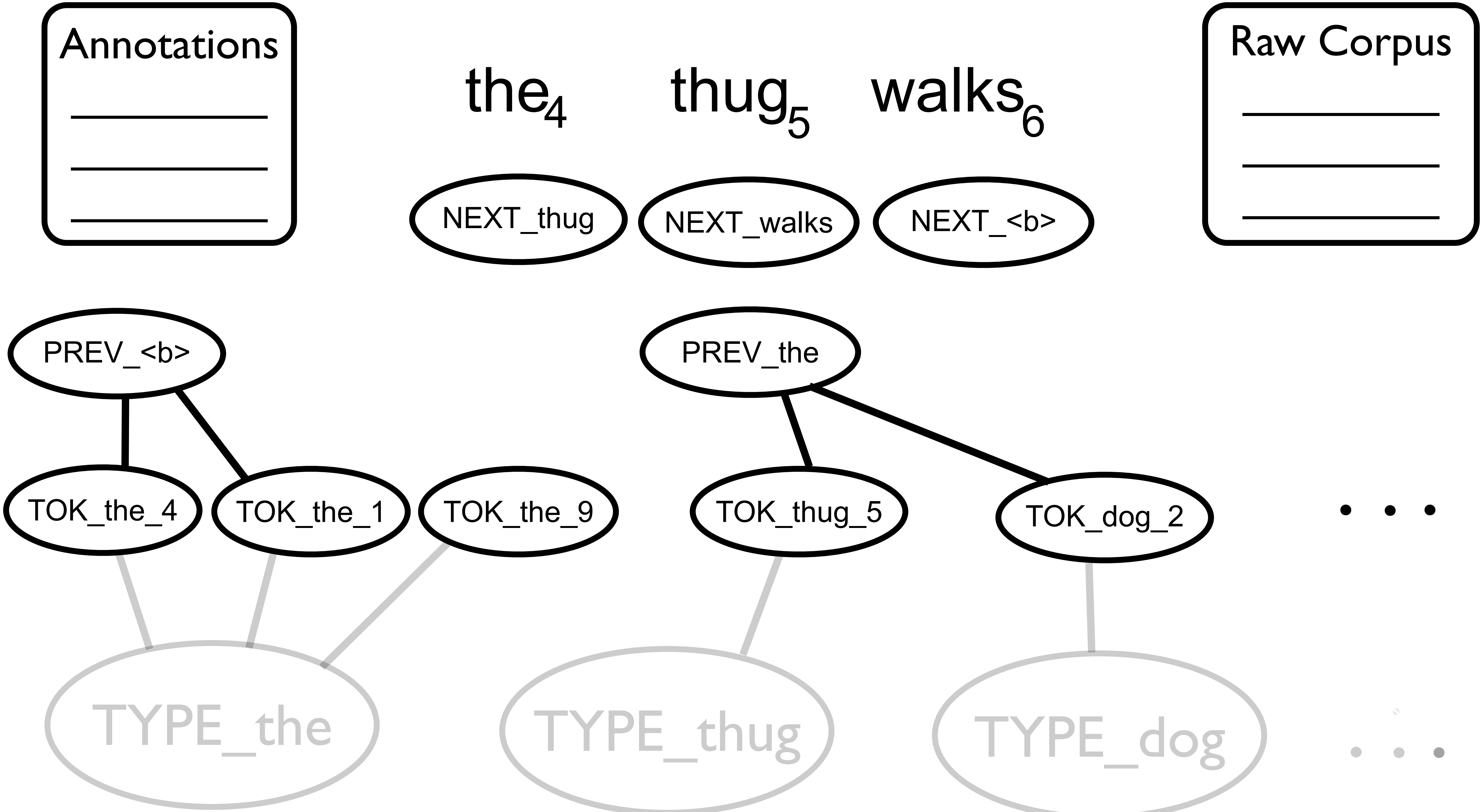
Annotations

the_4 thug_5 walks_6

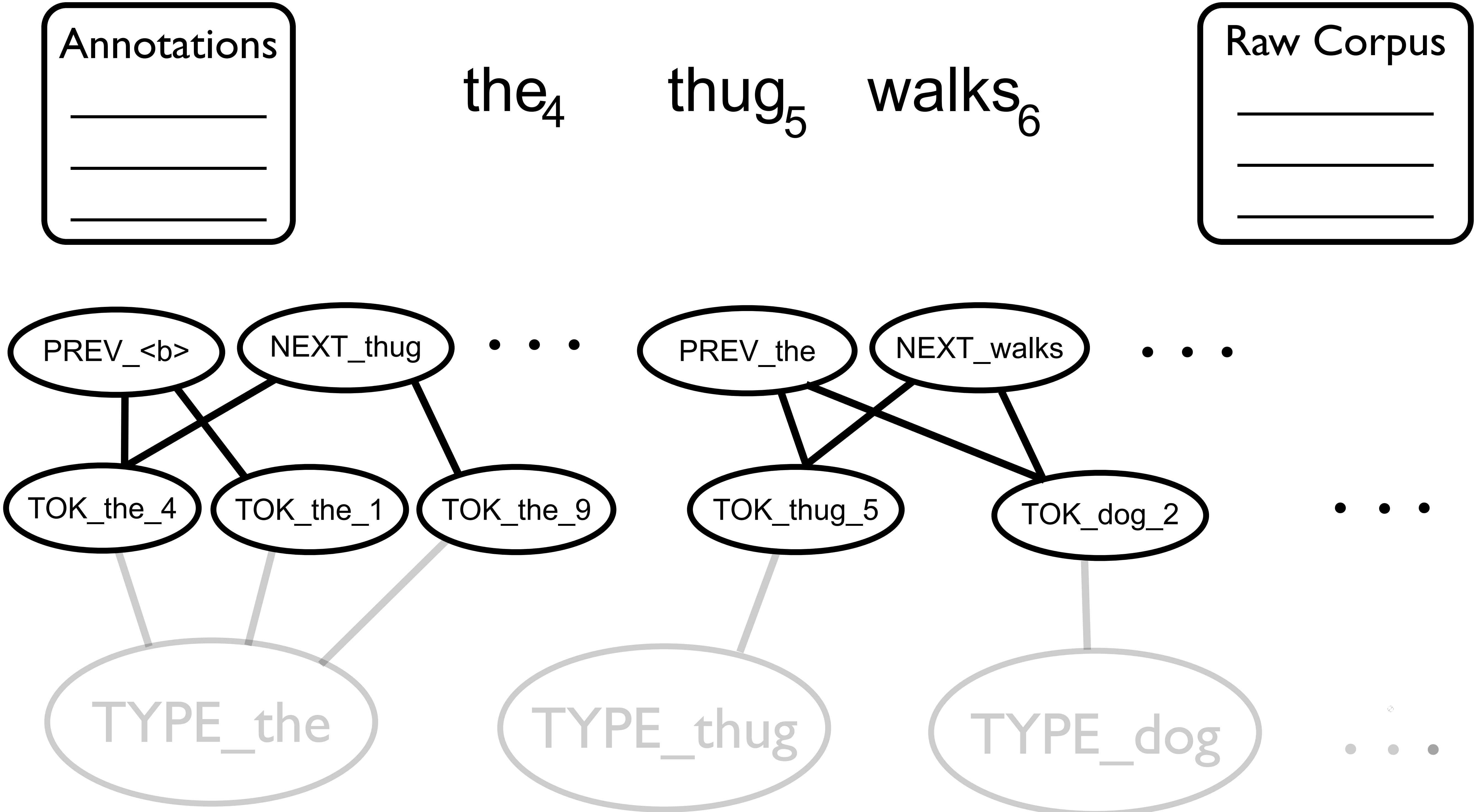
Raw Corpus



Tag Dict Generalization



Tag Dict Generalization



Annotations

the$_4$ thug$_5$ walks$_6$

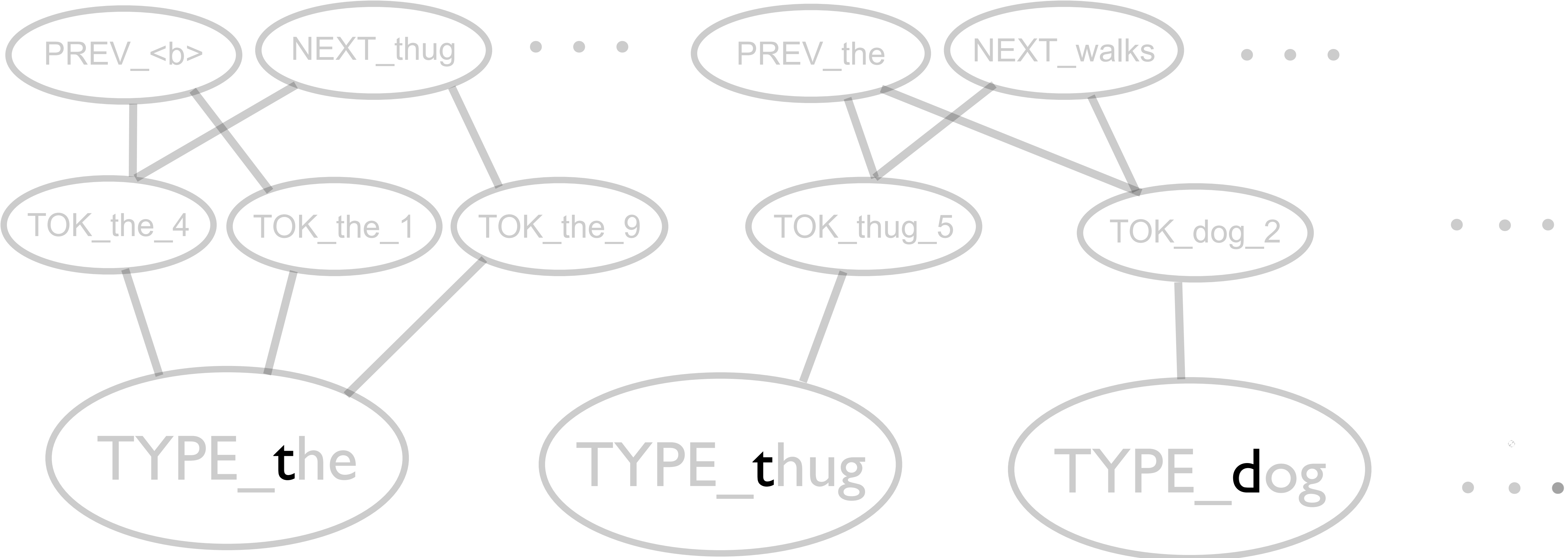
Raw Corpus

PREV_<b>
NEXT_thug
. . .
PREV_the
NEXT_walks
. . .

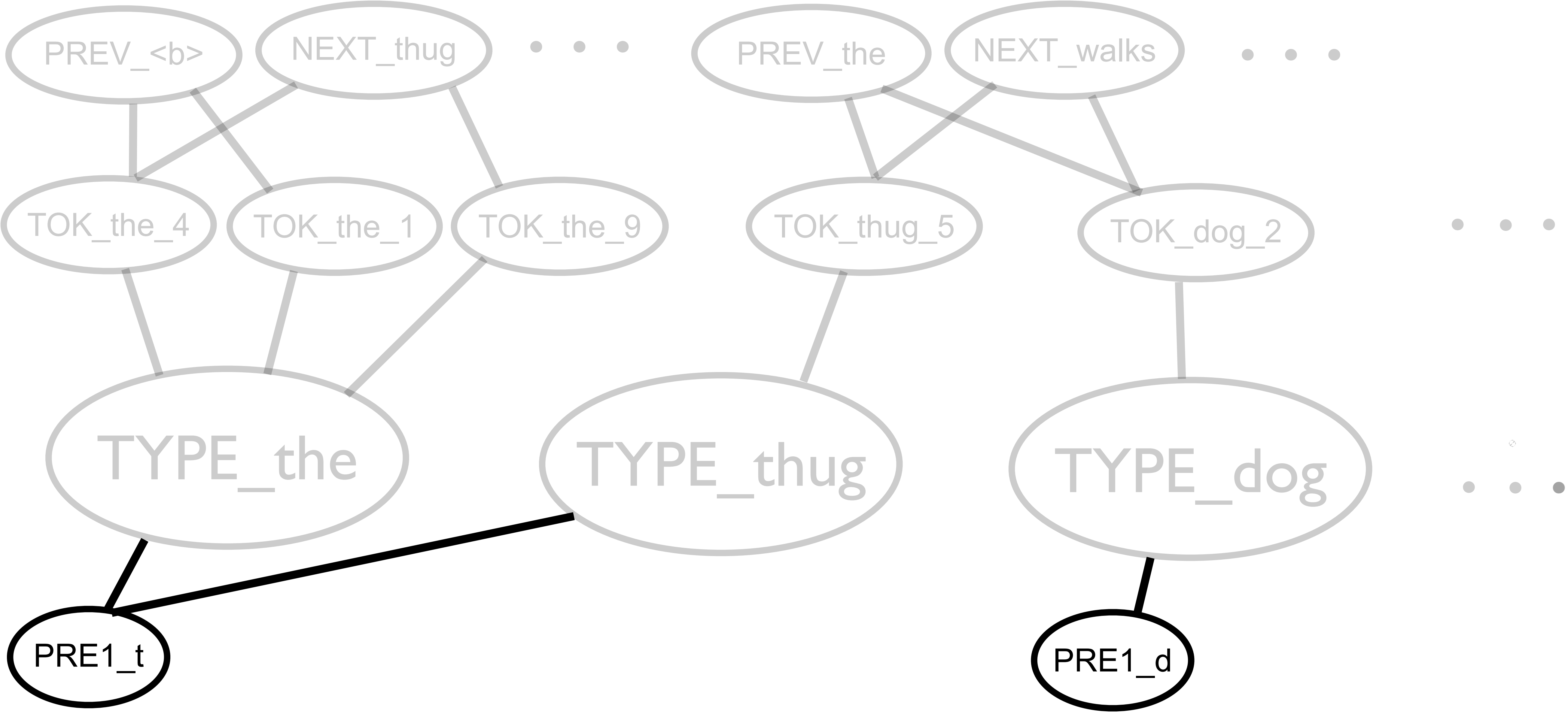
TOK_the_4
TOK_the_1
TOK_the_9
TOK_thug_5
TOK_dog_2
. . .

TYPE_the
TYPE_thug
TYPE_dog
. . .

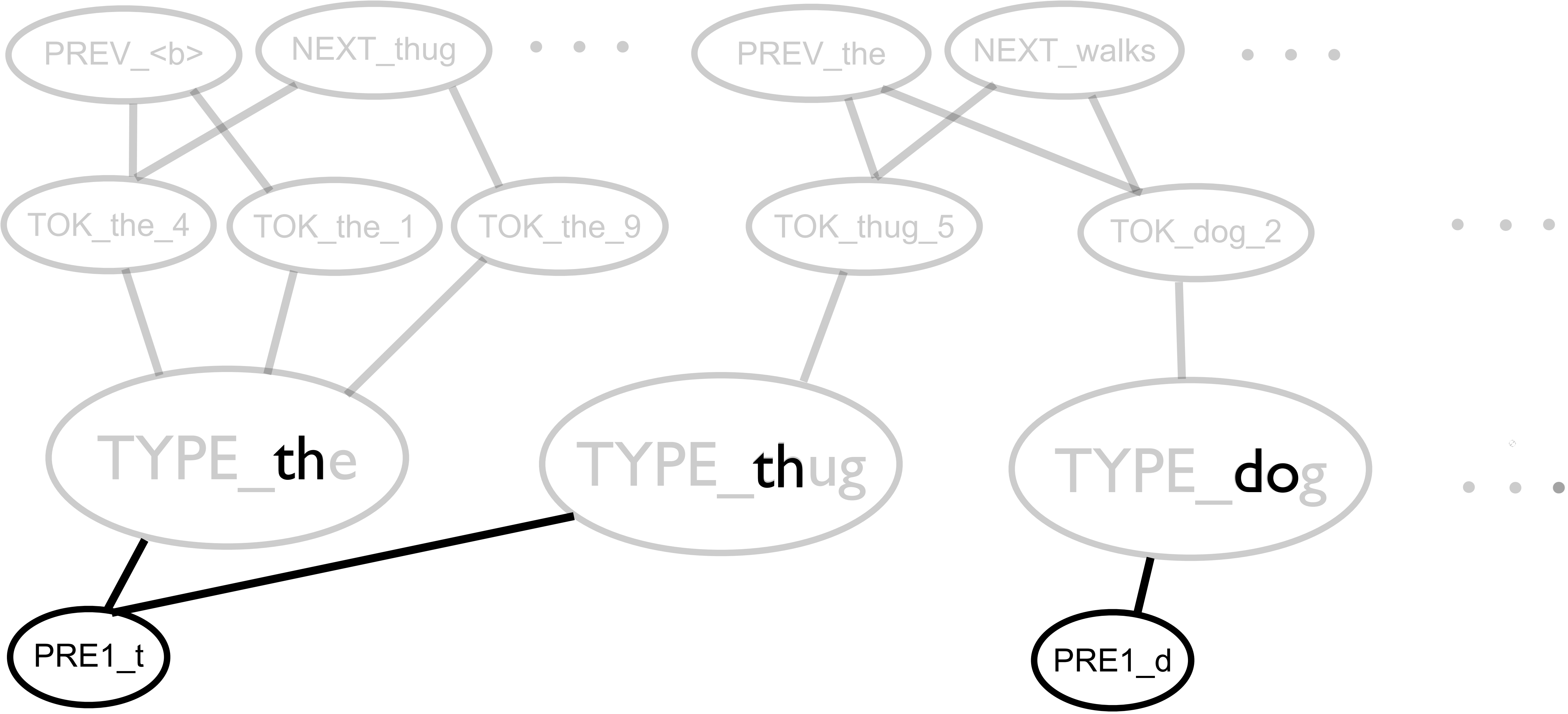
Tag Dict Generalization



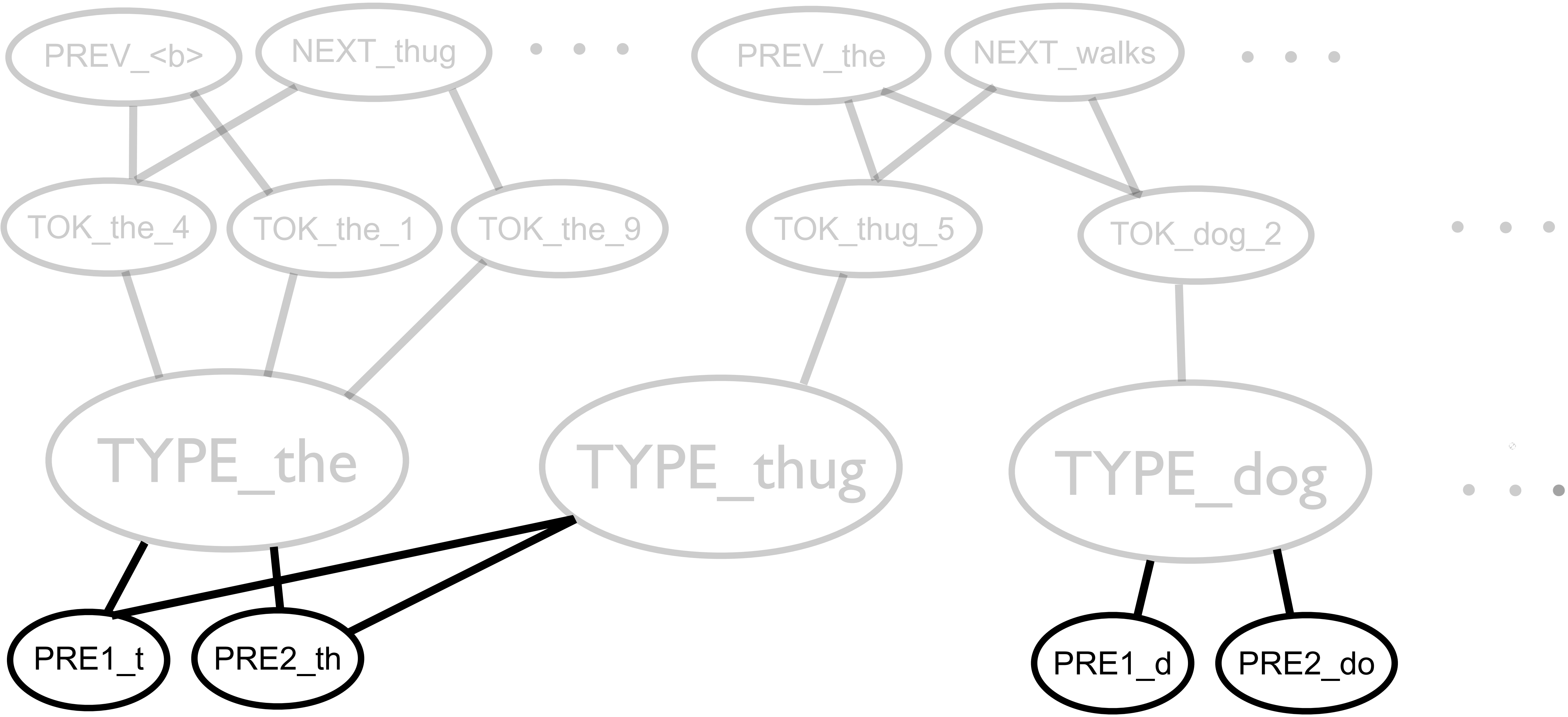
Tag Dict Generalization



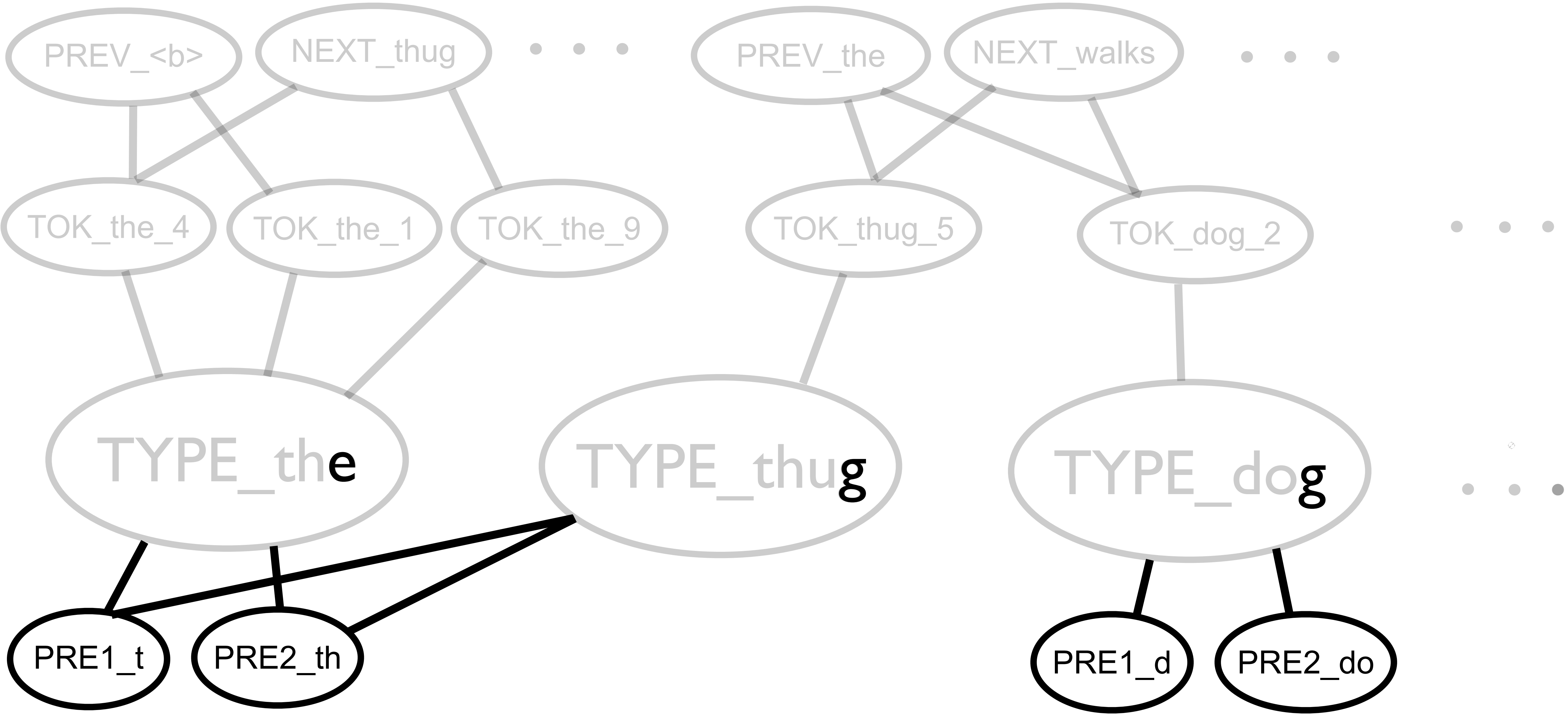
Tag Dict Generalization



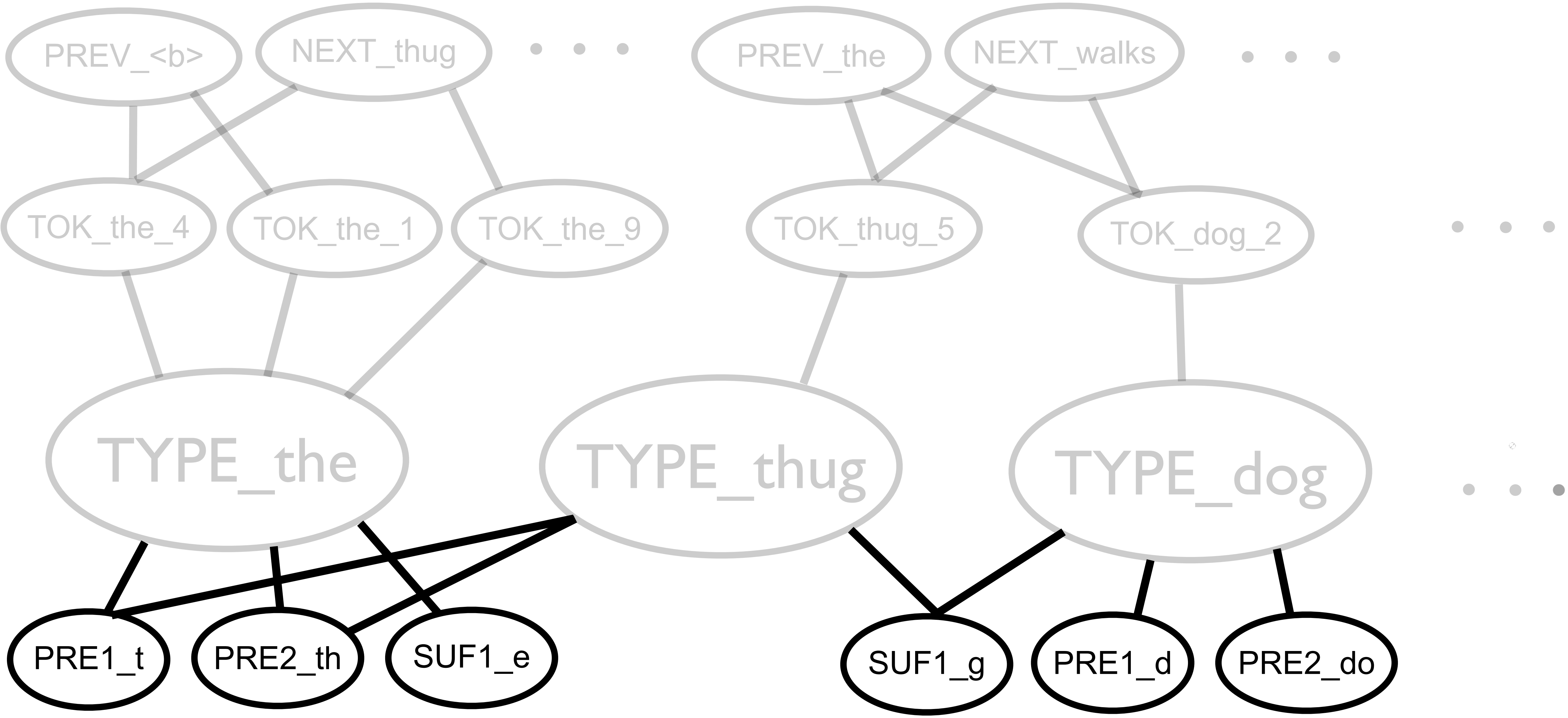
Tag Dict Generalization



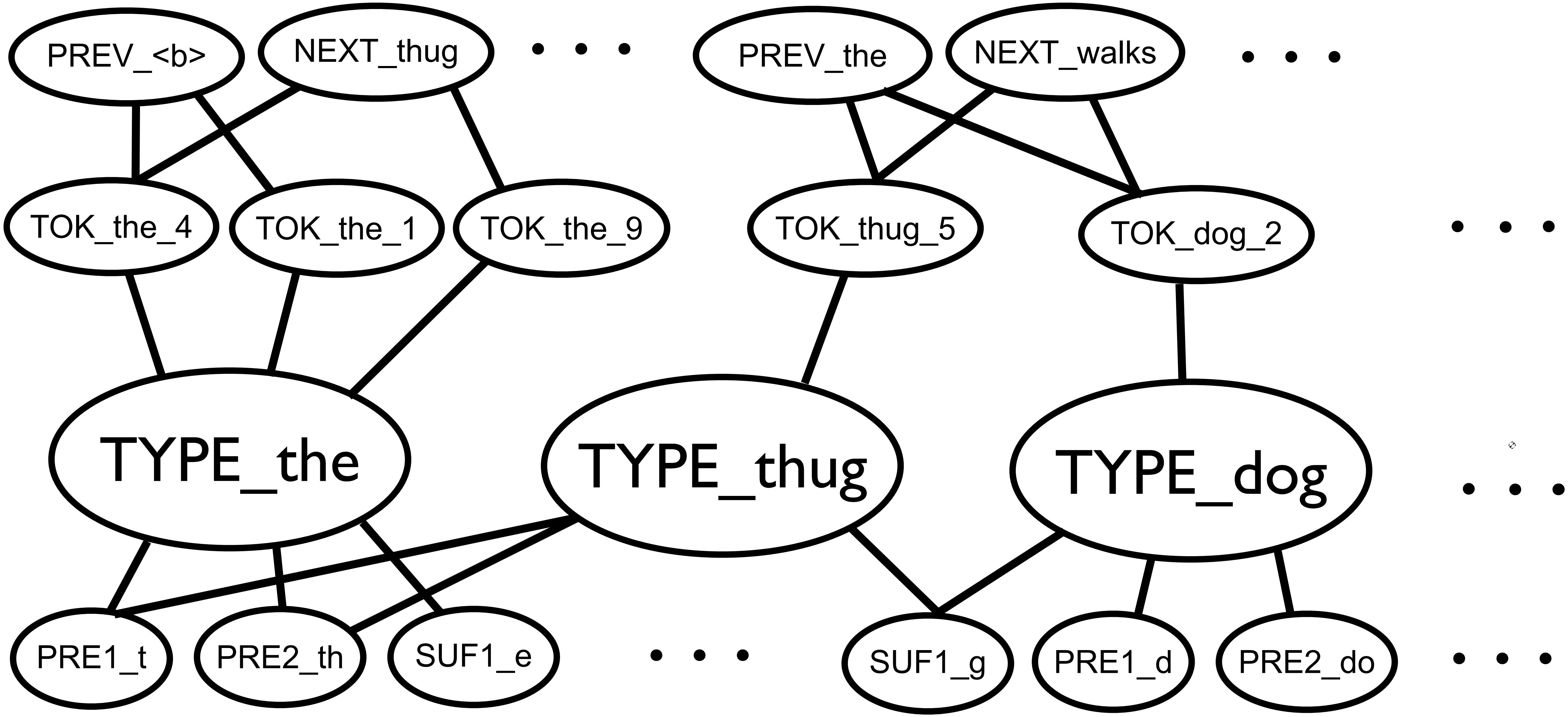
Tag Dict Generalization



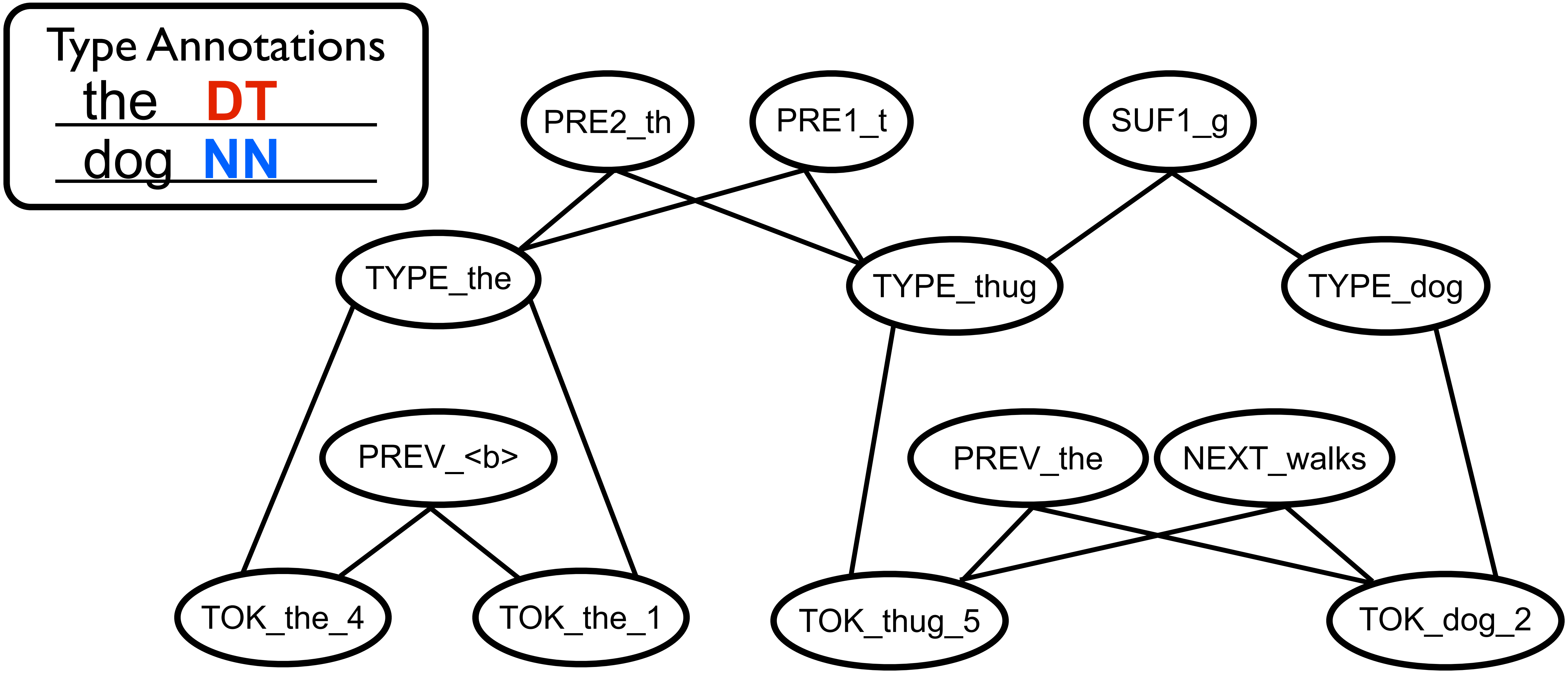
Tag Dict Generalization



Tag Dict Generalization



Tag Dict Generalization



Type Annotations

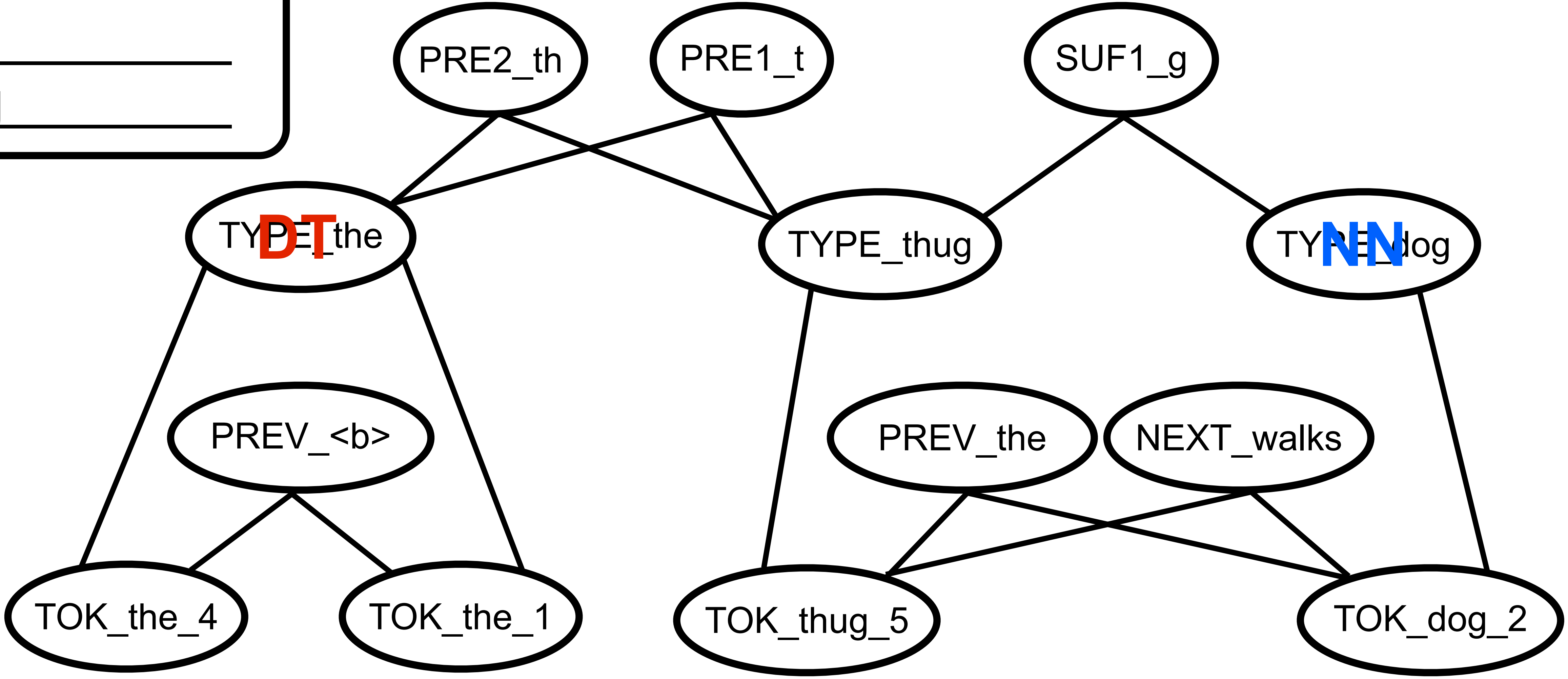
the	DT
dog	NN

Tag Dict Generalization

Type Annotations

the

dog

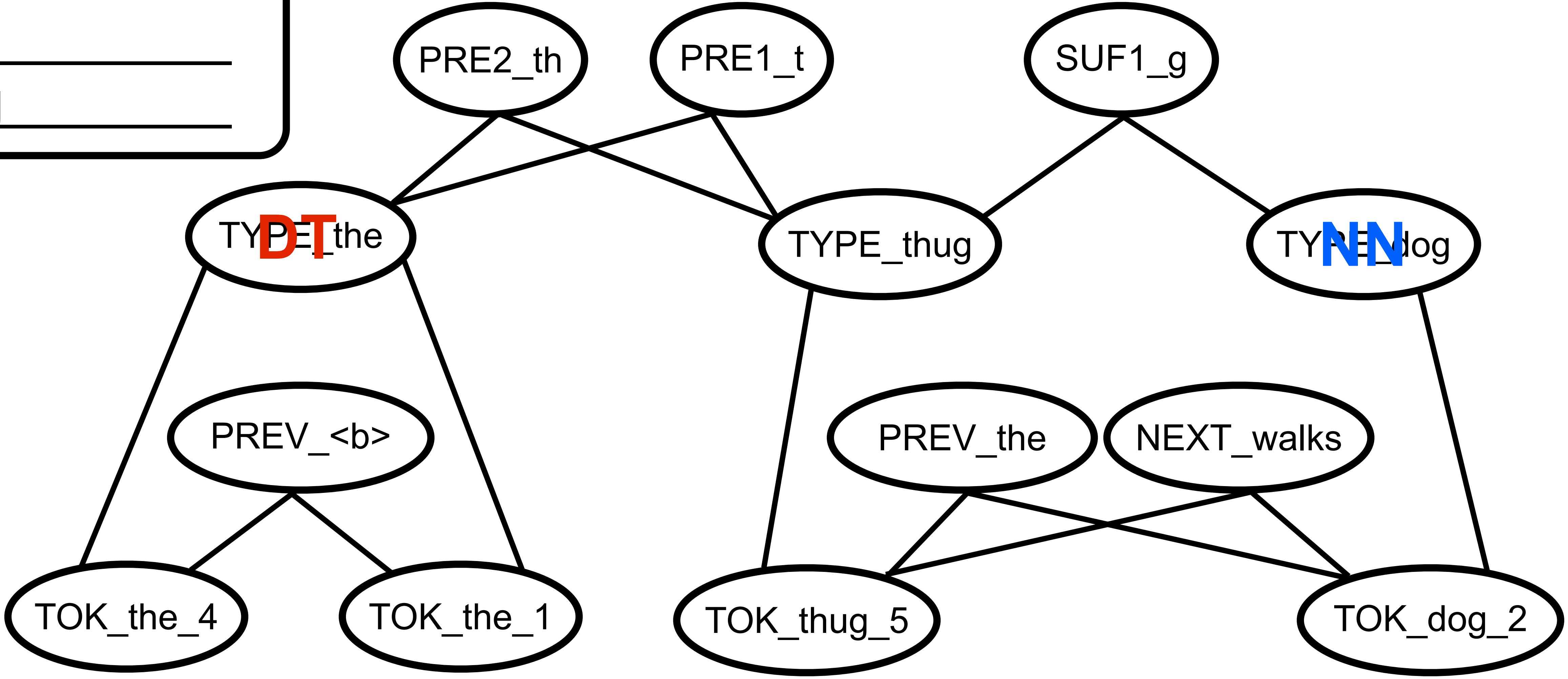


Tag Dict Generalization

Type Annotations

the

dog



Token Annotations

the dog walks

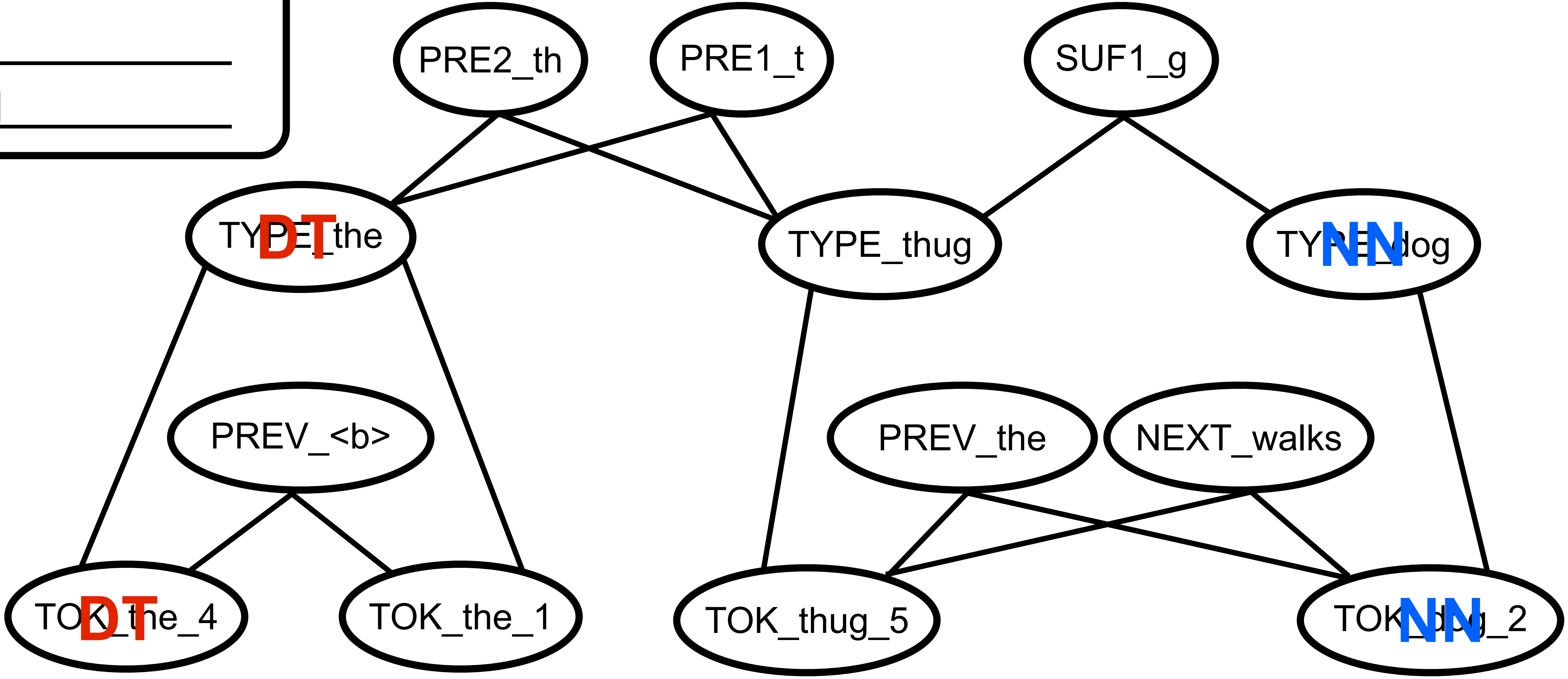
DT NN VBZ

Tag Dict Generalization

Type Annotations

the

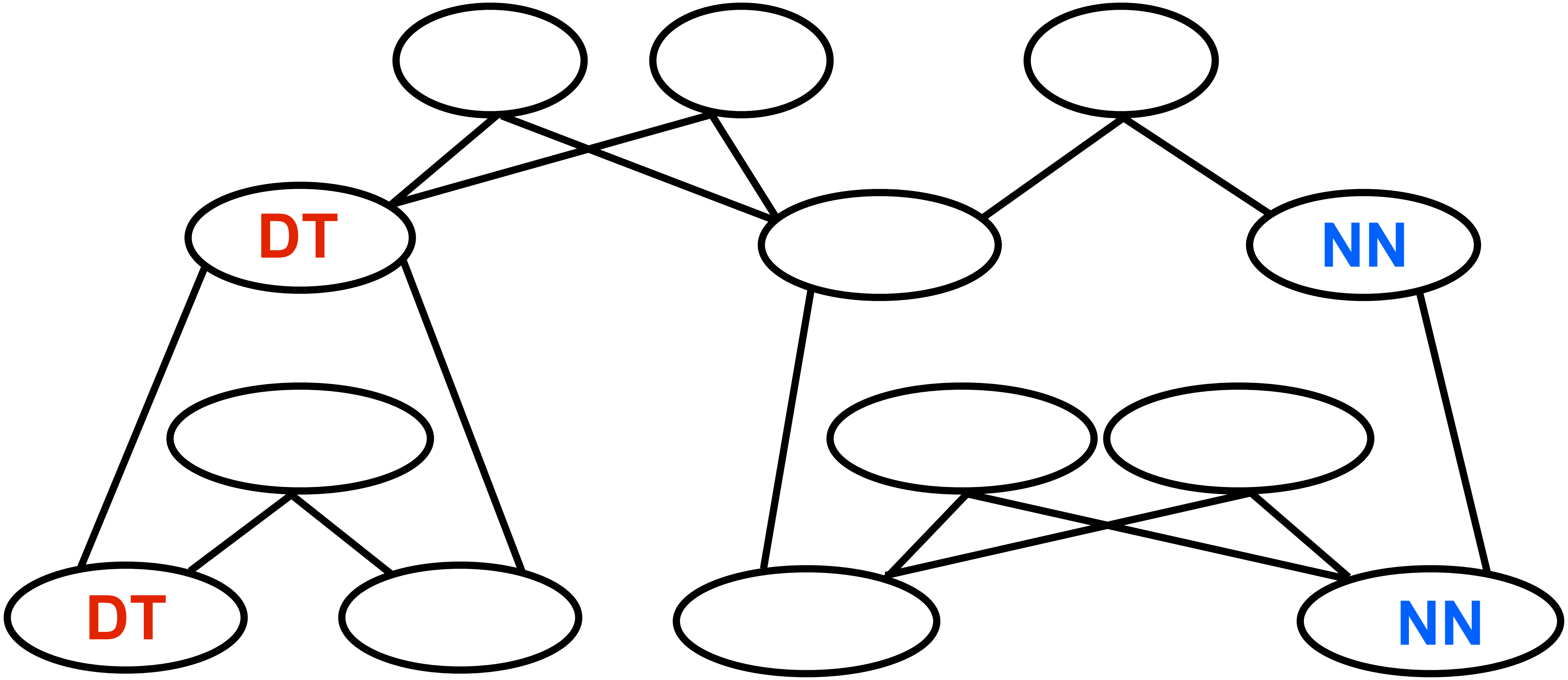
dog



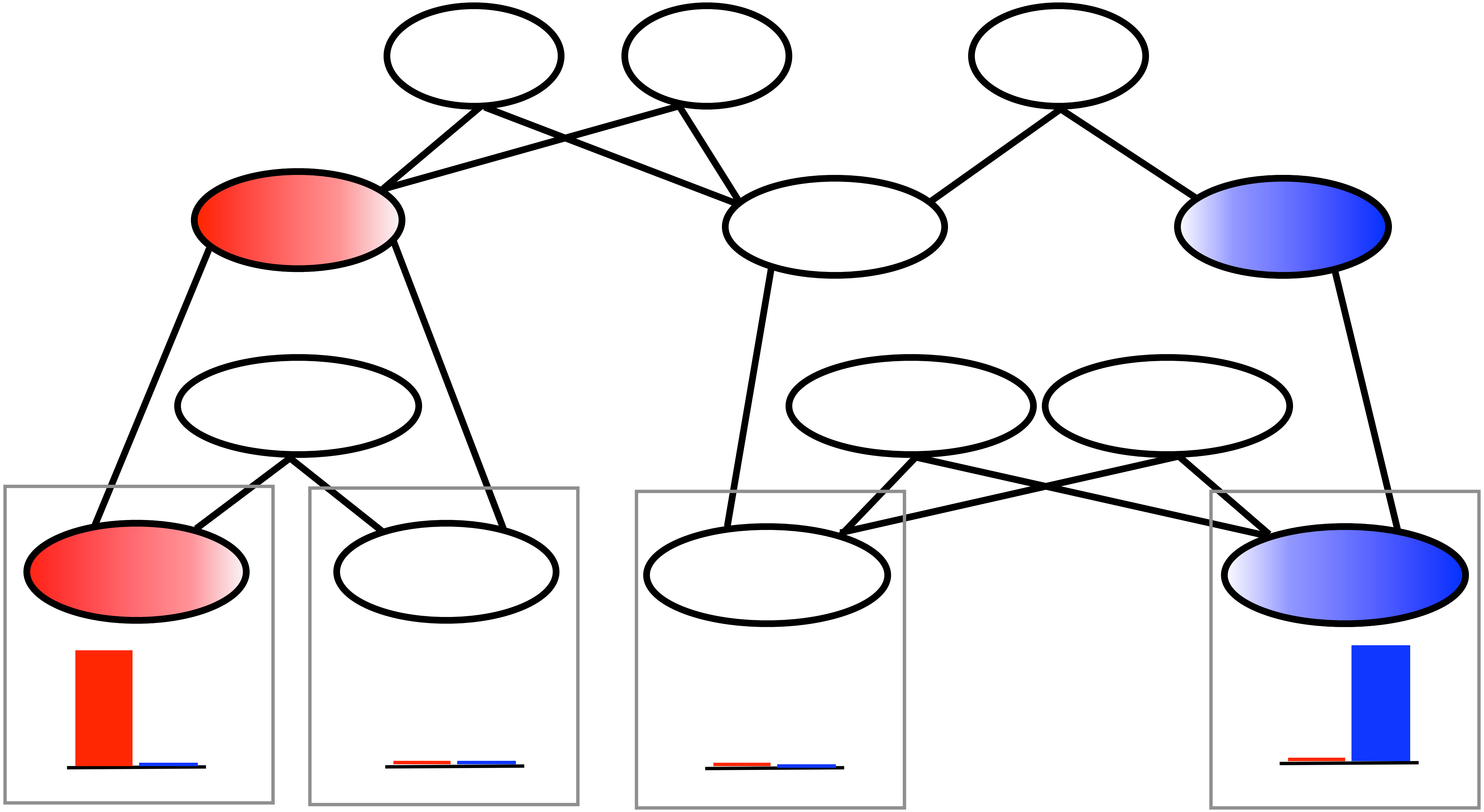
Token Annotations

the dog walks

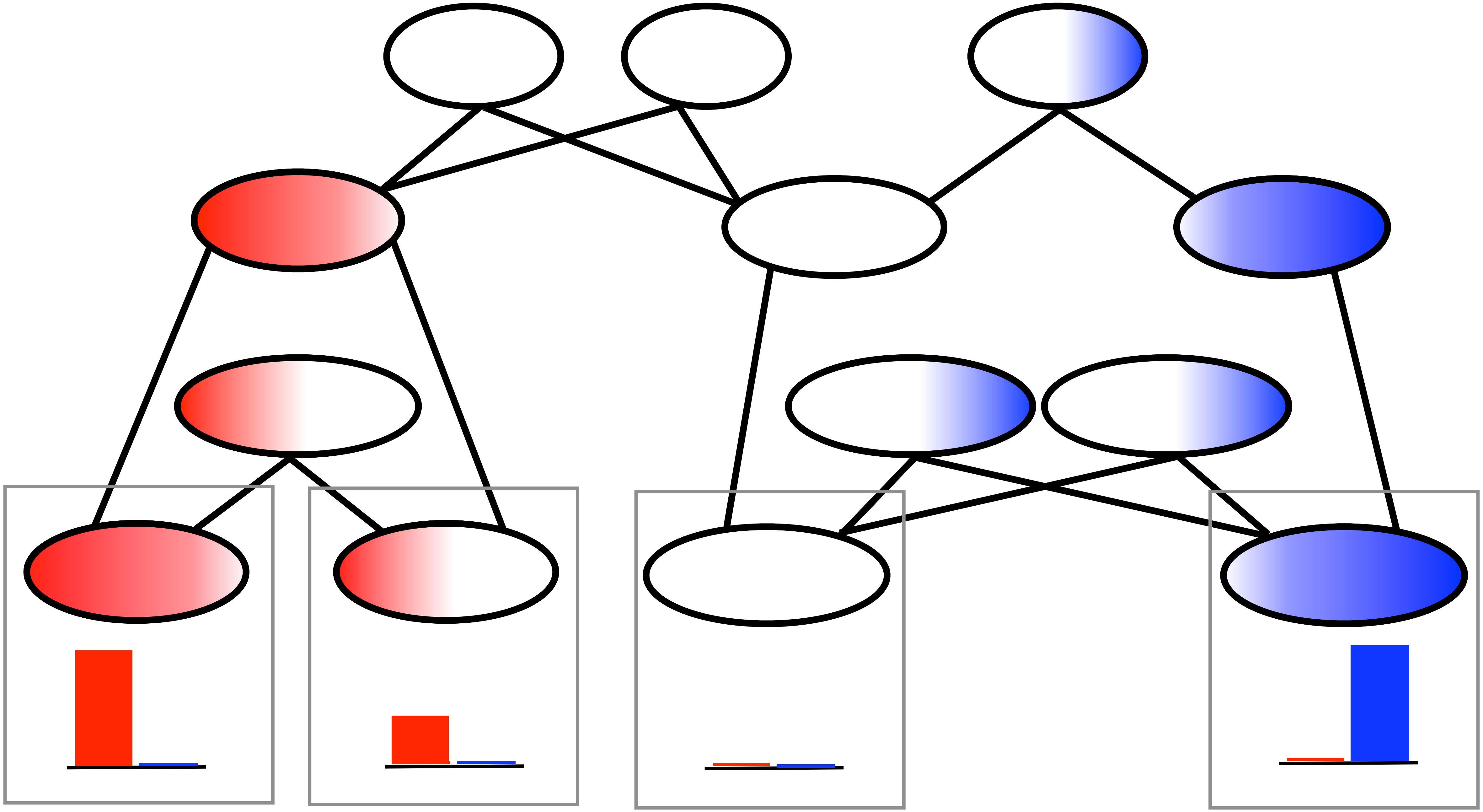
Tag Dict Generalization



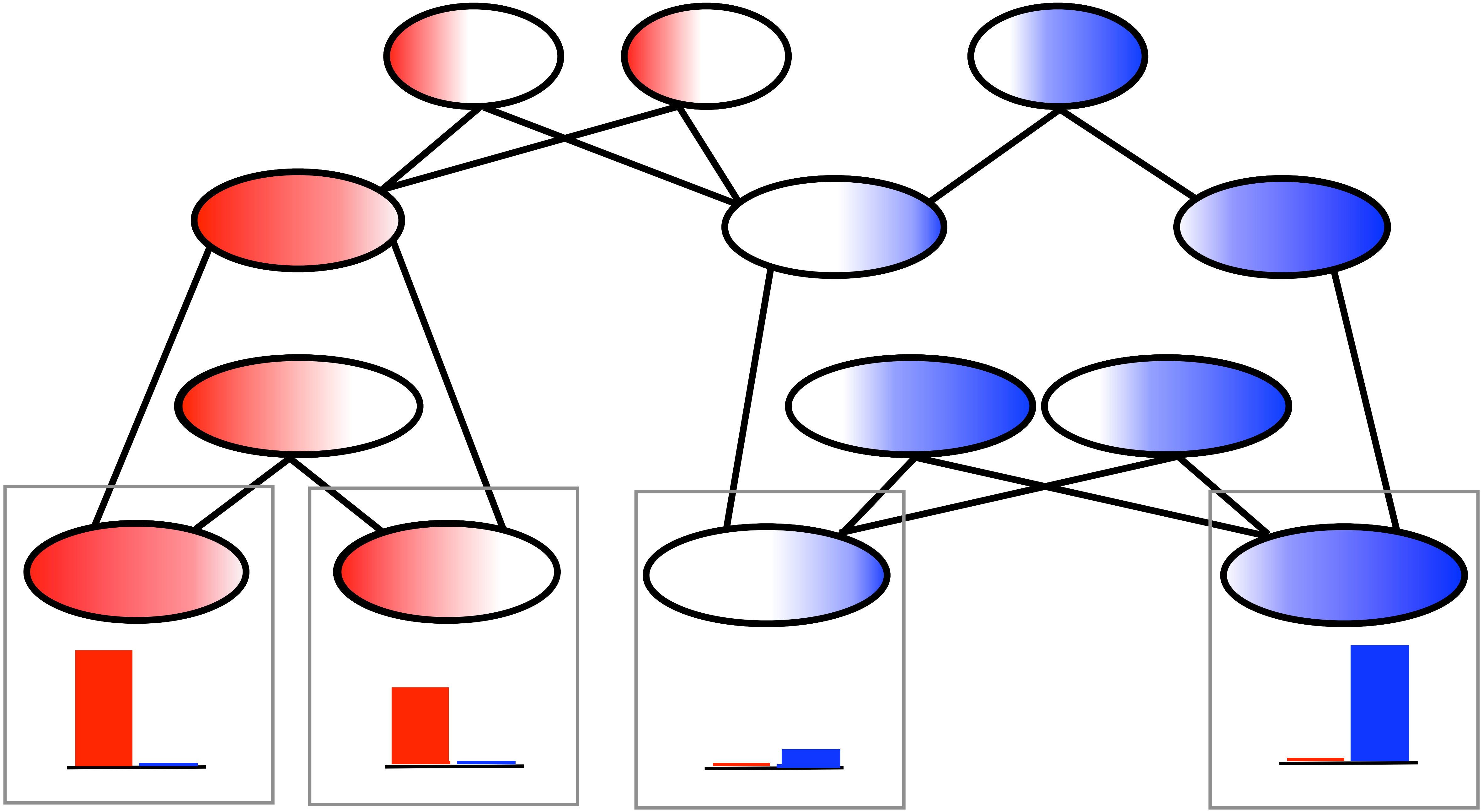
Tag Dict Generalization



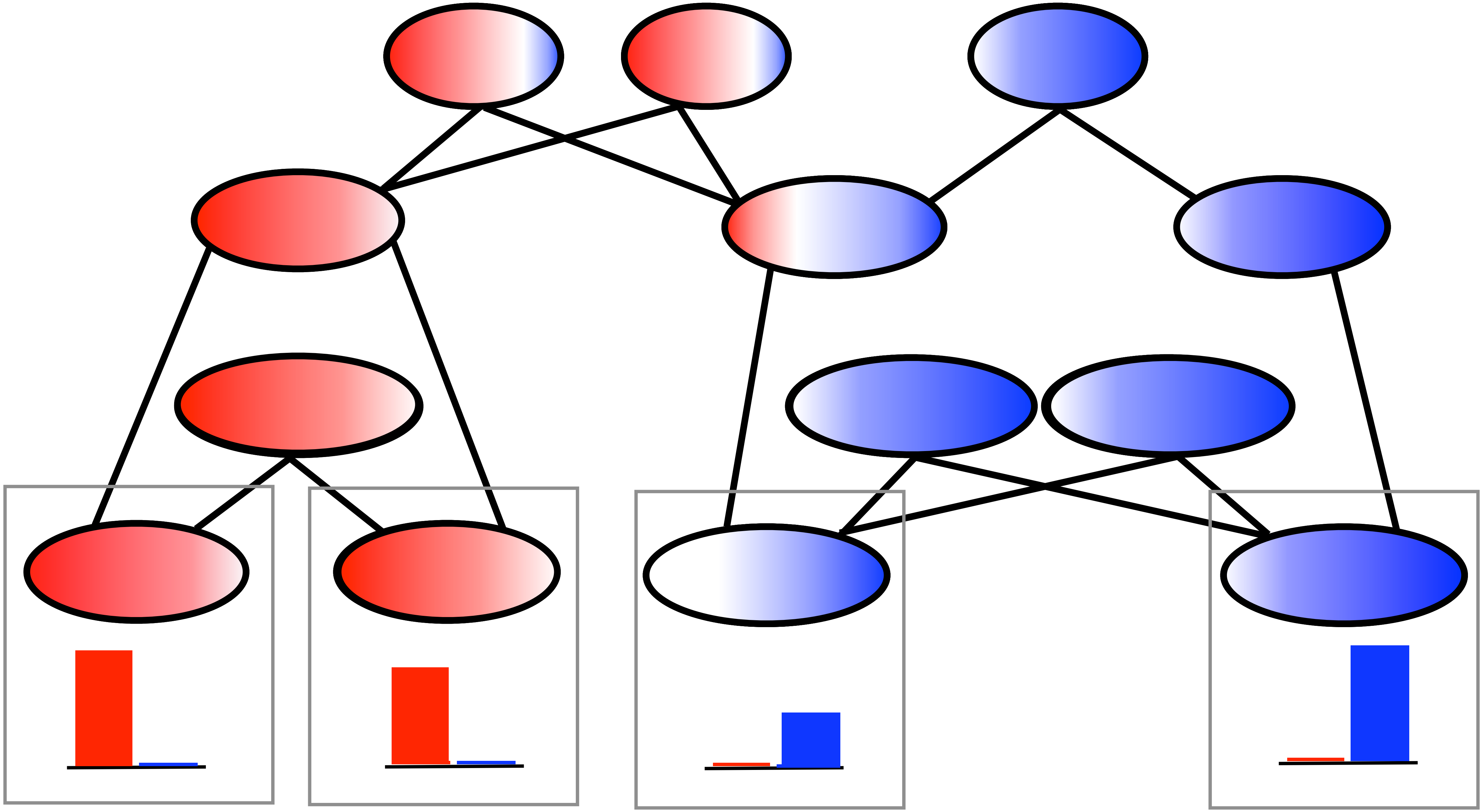
Tag Dict Generalization



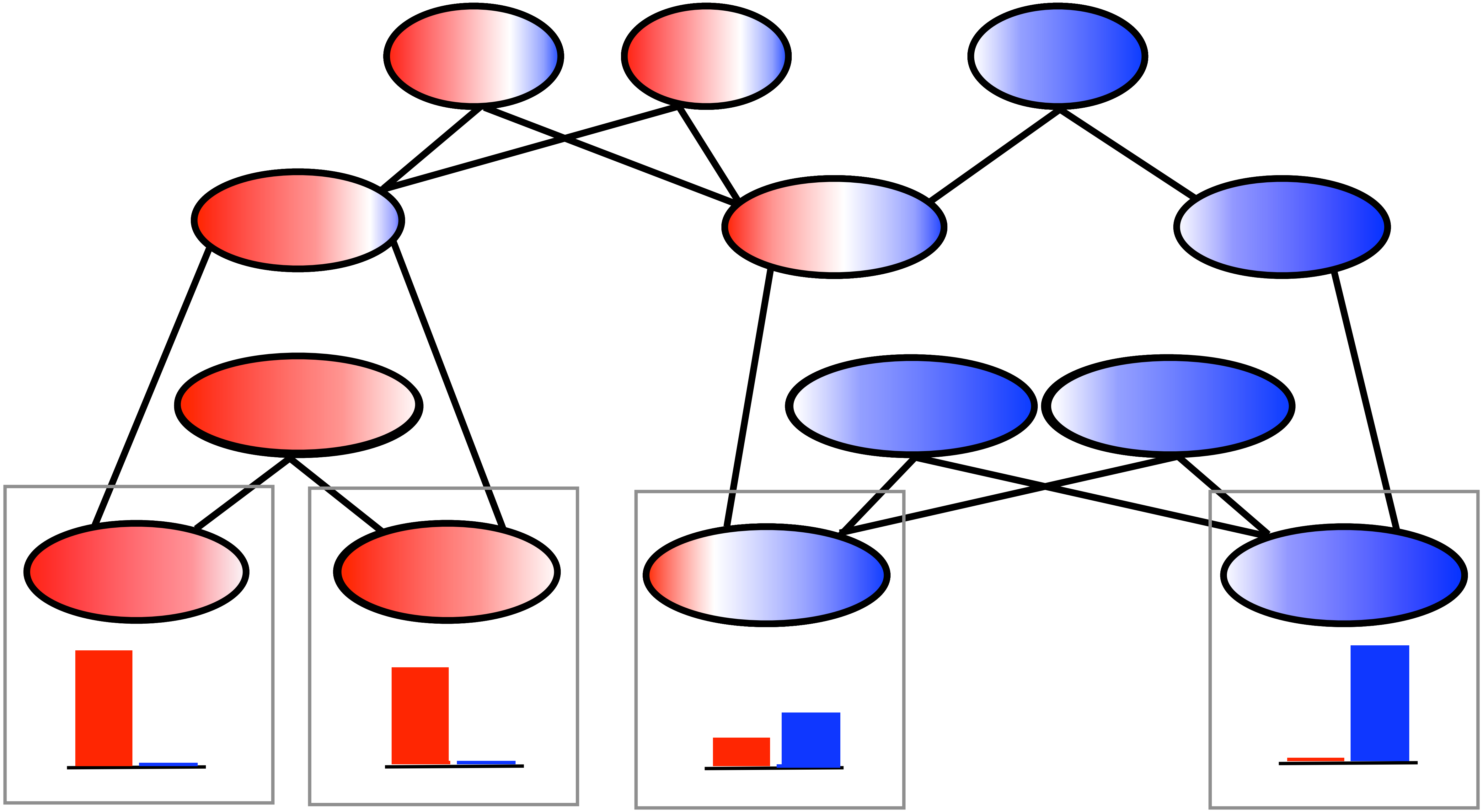
Tag Dict Generalization



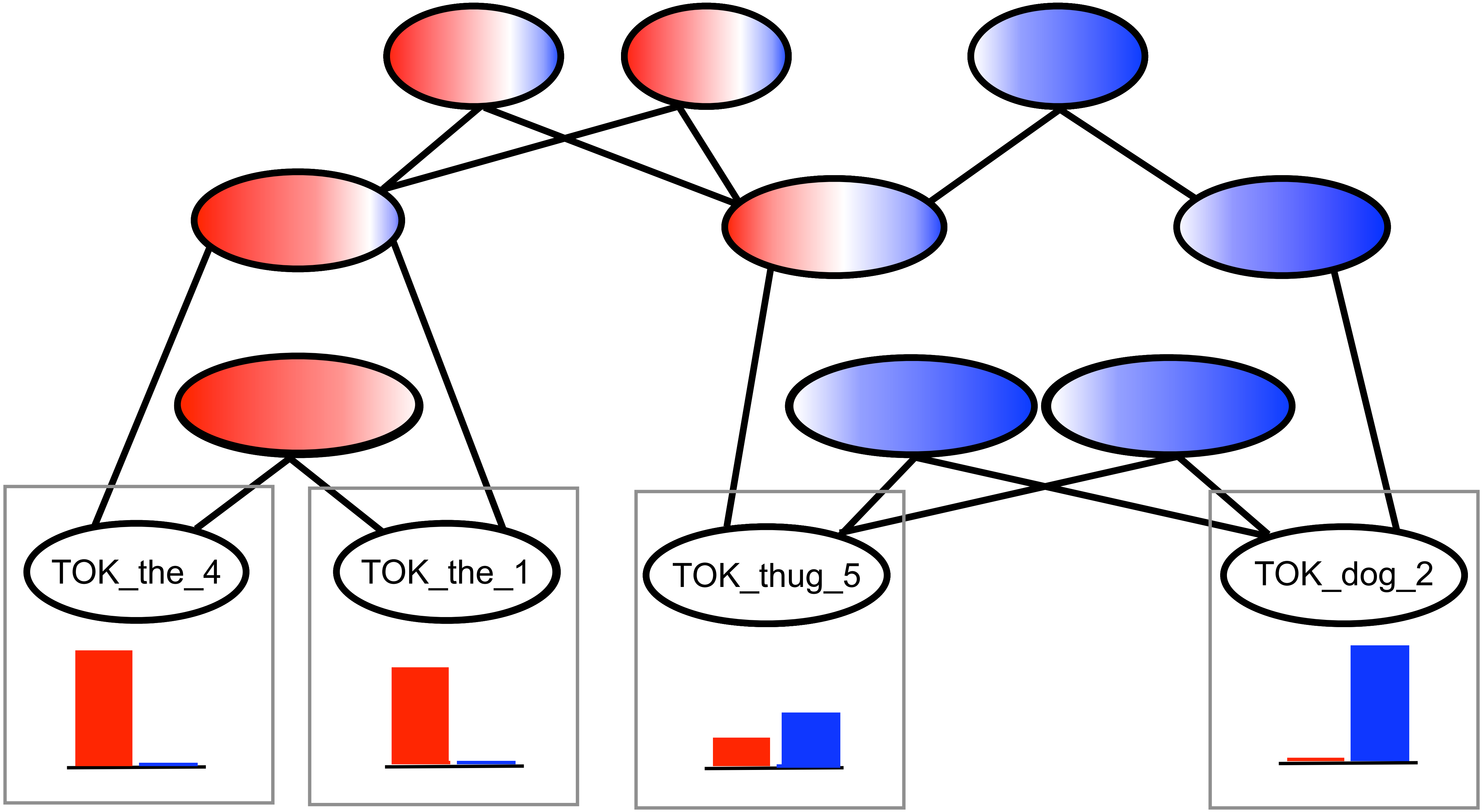
Tag Dict Generalization



Tag Dict Generalization



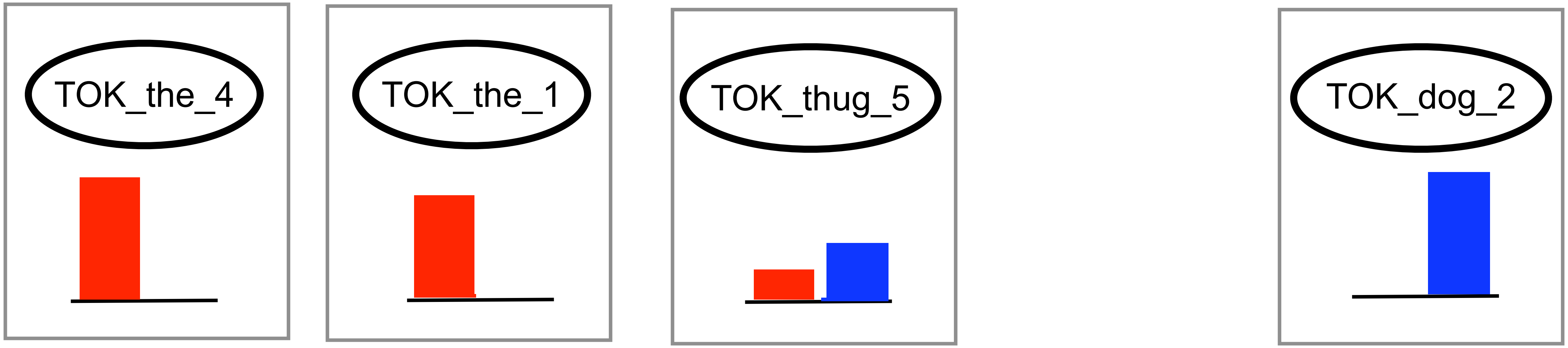
Tag Dict Generalization



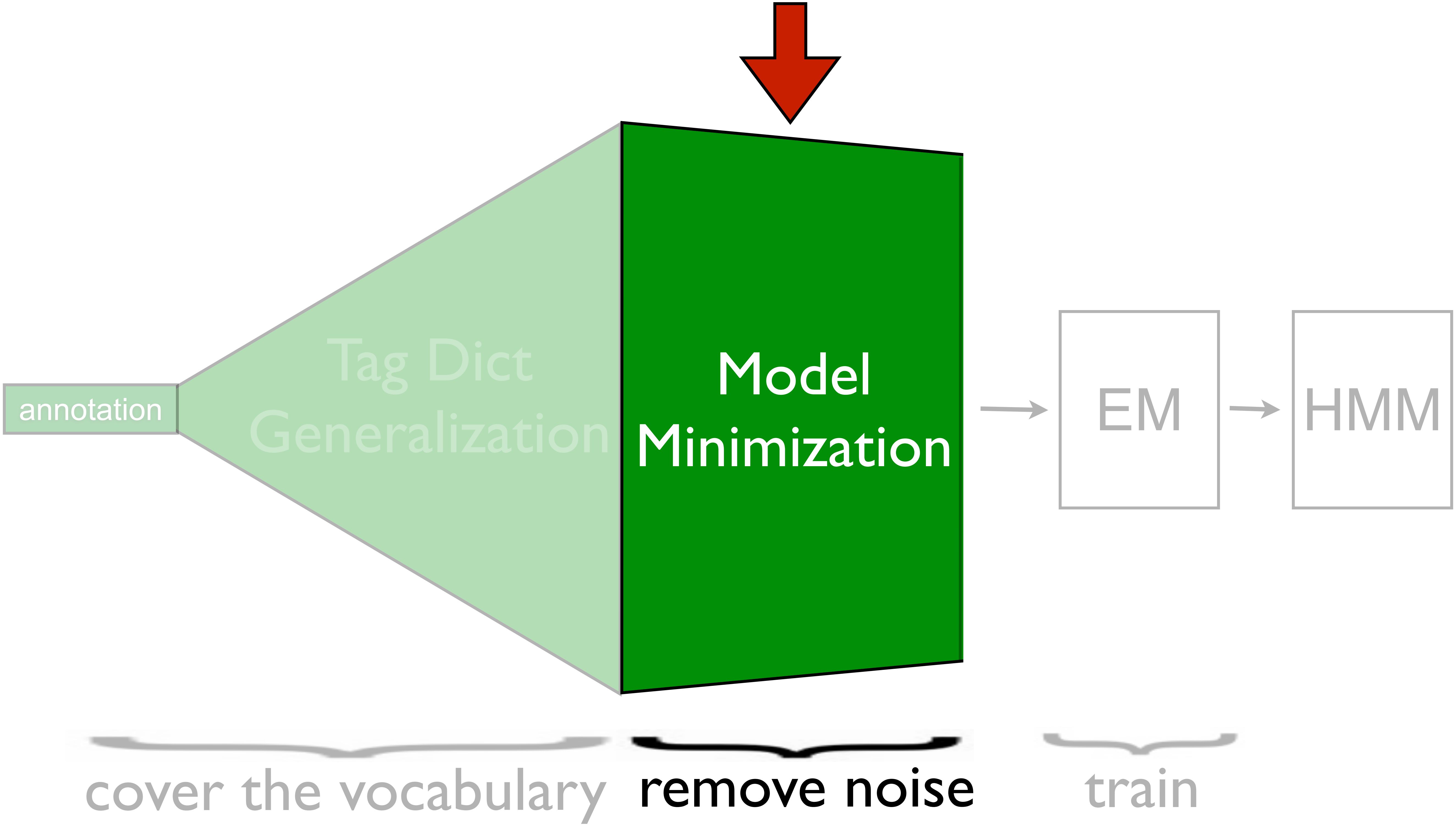
Tag Dict Generalization

Result:

- a tag distribution on every token (soft tagging)
- an expanded tag dictionary (non-zero tags)



Our Approach

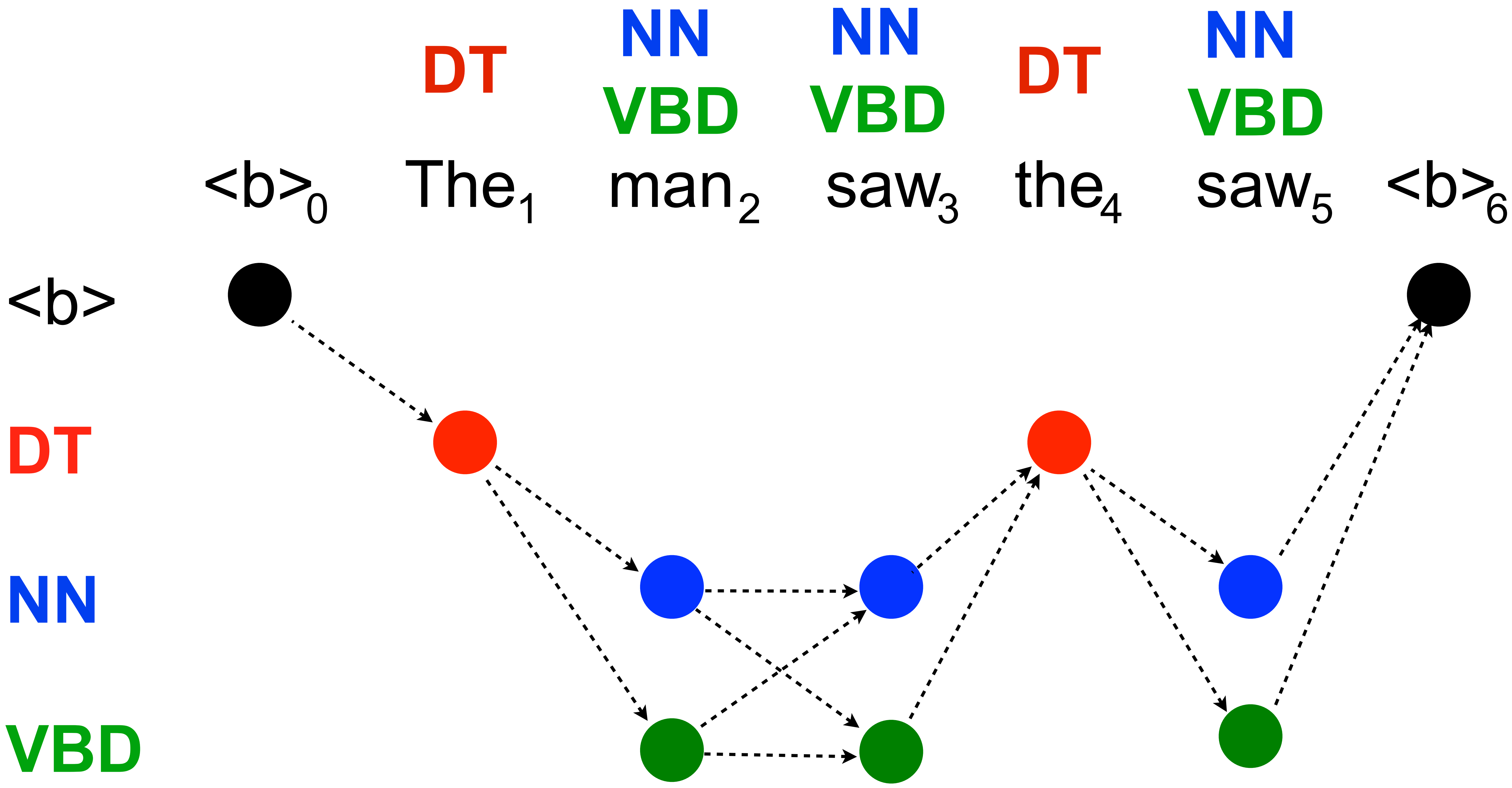


Model Minimization

- Induce a cleaner hard tagging from a noisy soft tagging.
- Greedily seek the minimal set of tag bigrams that describe the raw corpus.

[Ravi et al., 2010; Garrette and Baldridge, 2012]

Model Minimization



DT NN VBD NN VBD DT NN VBD

<b>$_0$ The$_1$ man$_2$ saw$_3$ the$_4$ saw$_5$ <b>$_6$

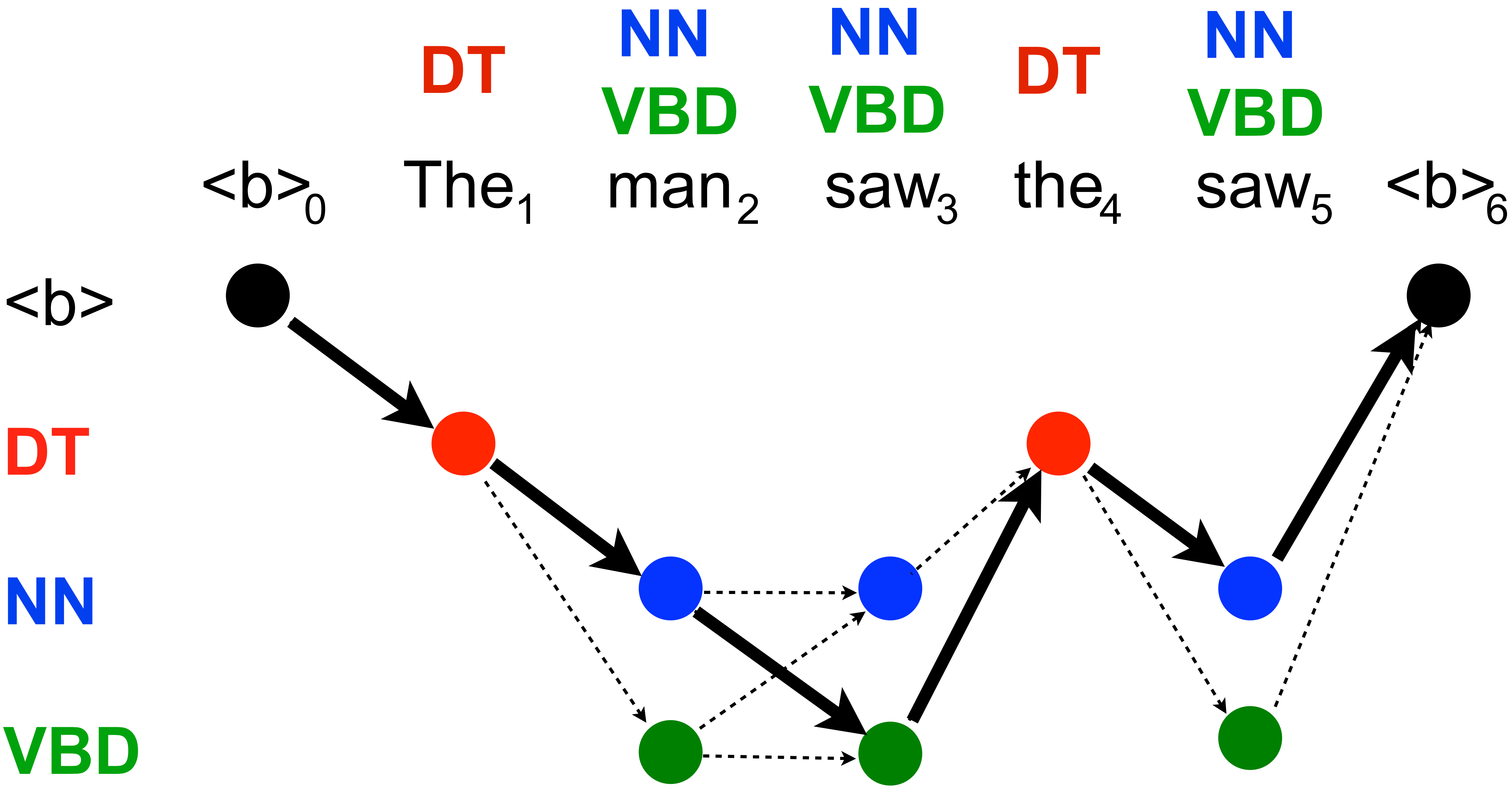
<b>

DT

NN

VBD

Model Minimization



DT NN VBD NN VBD DT NN VBD

<b>$_0$ The$_1$ man$_2$ saw$_3$ the$_4$ saw$_5$ <b>$_6$

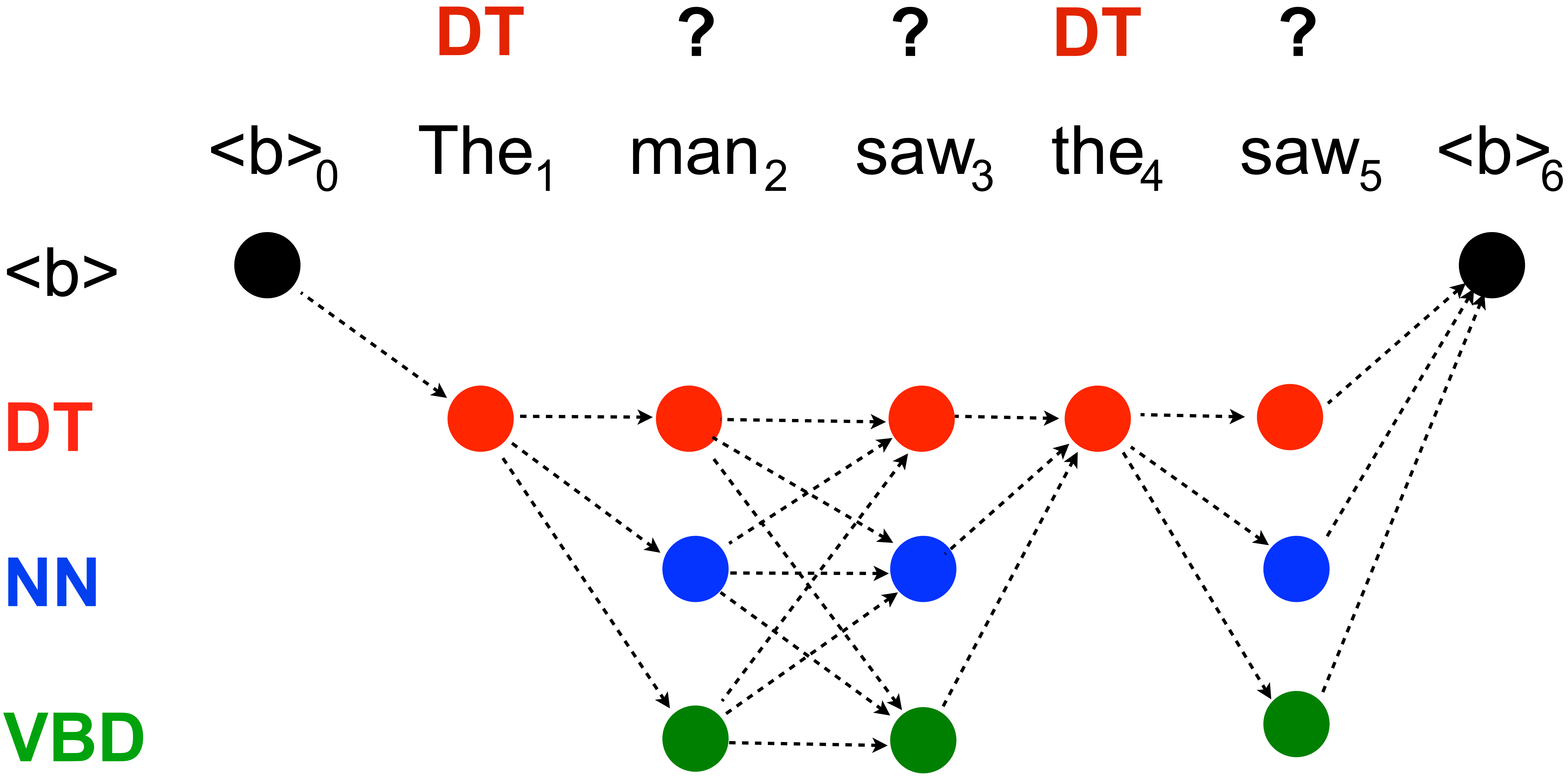
<b>

DT

NN

VBD

Model Minimization



DT ? ? DT ?

<b>$_0$ The$_1$ man$_2$ saw$_3$ the$_4$ saw$_5$ <b>$_6$

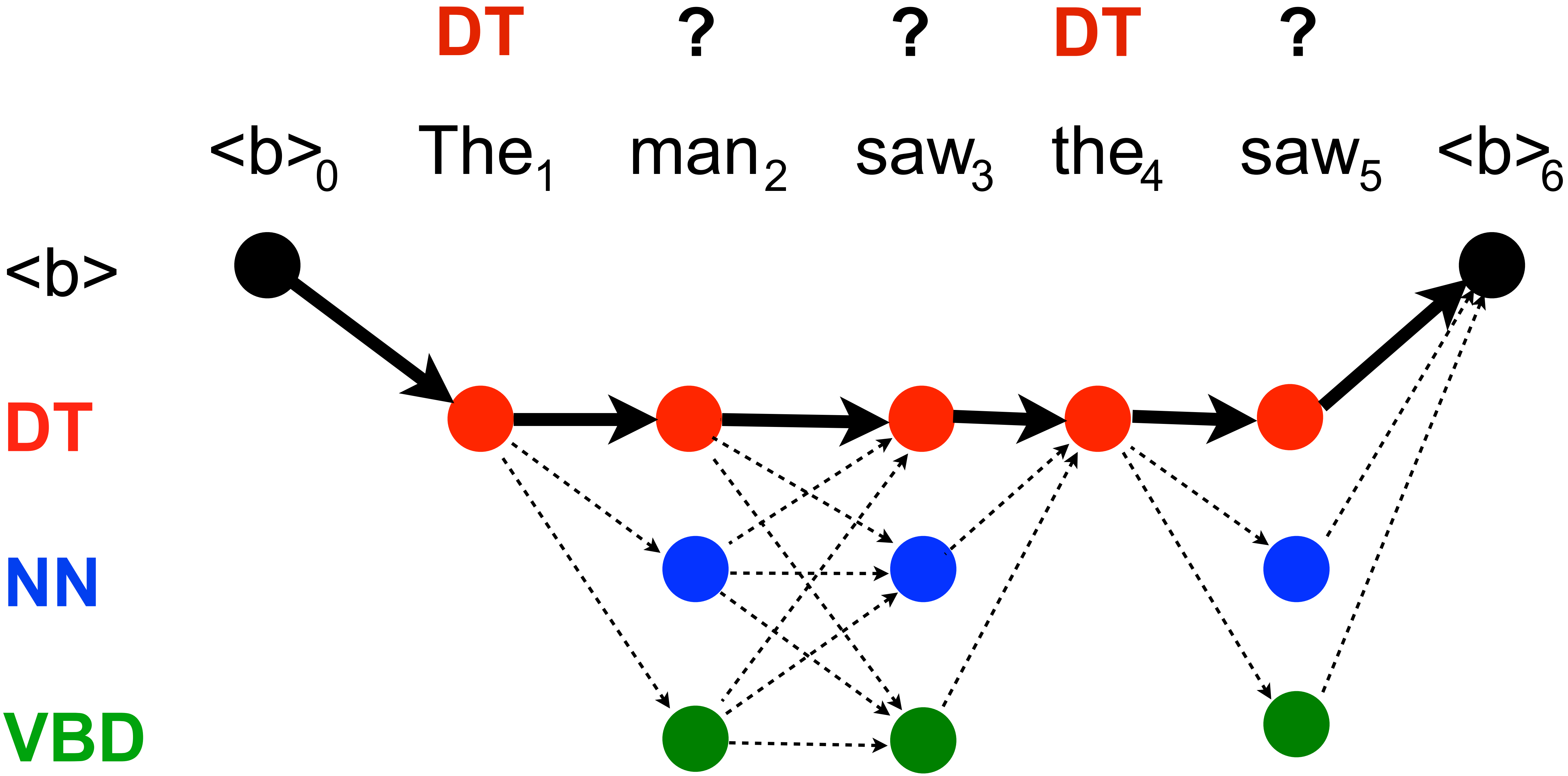
<b>

DT

NN

VBD

Model Minimization



DT ? ? DT ?

<b>$_0$ The$_1$ man$_2$ saw$_3$ the$_4$ saw$_5$ <b>$_6$

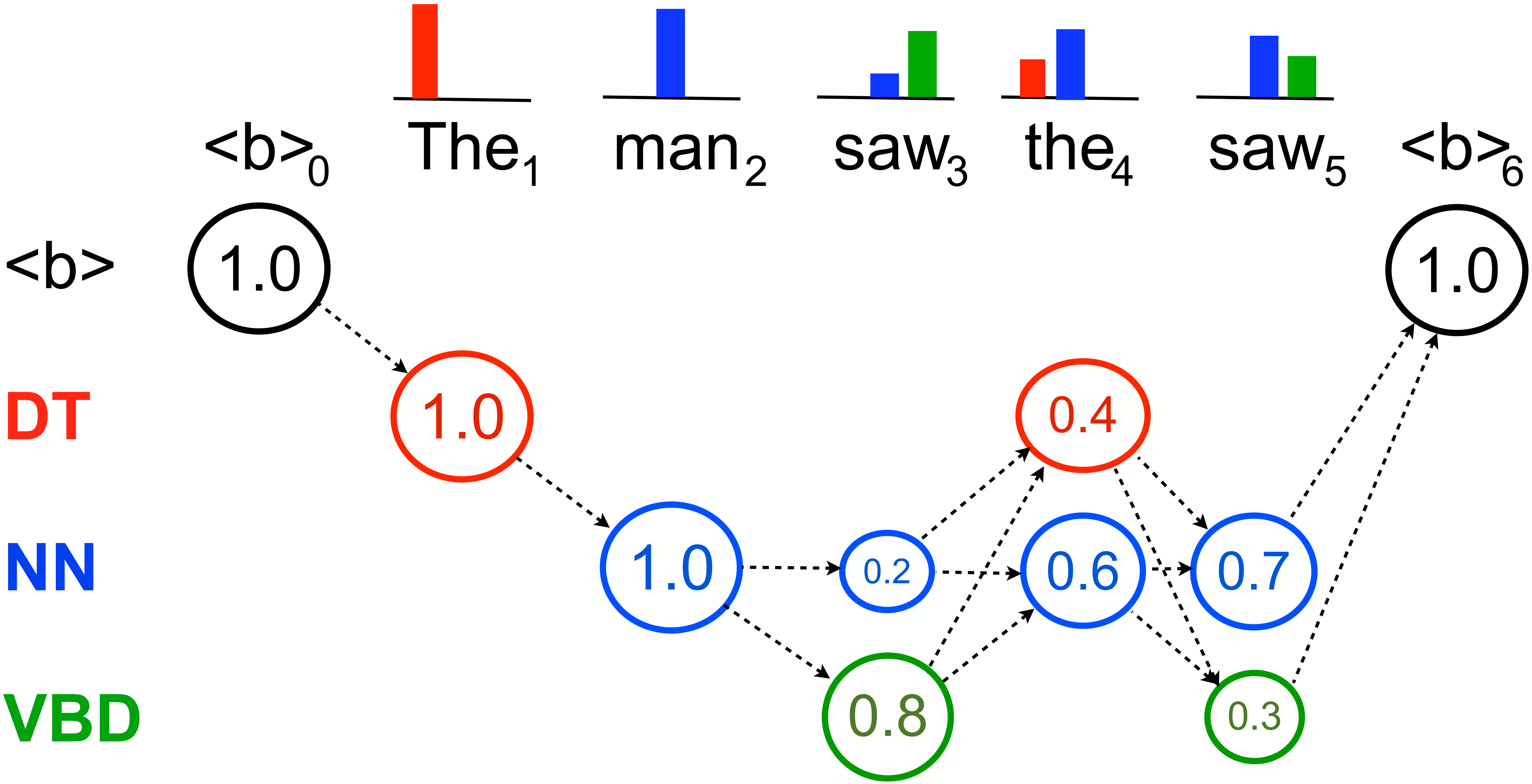
<b>

DT

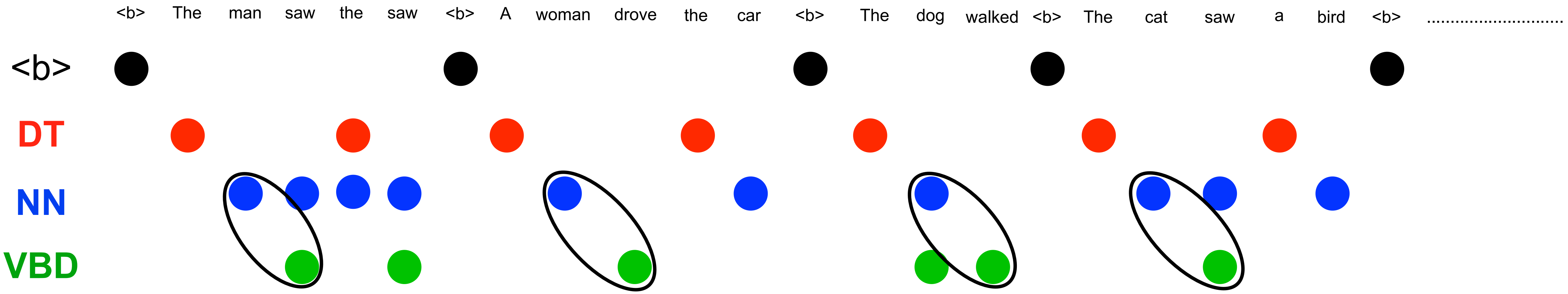
NN

VBD

Model Minimization

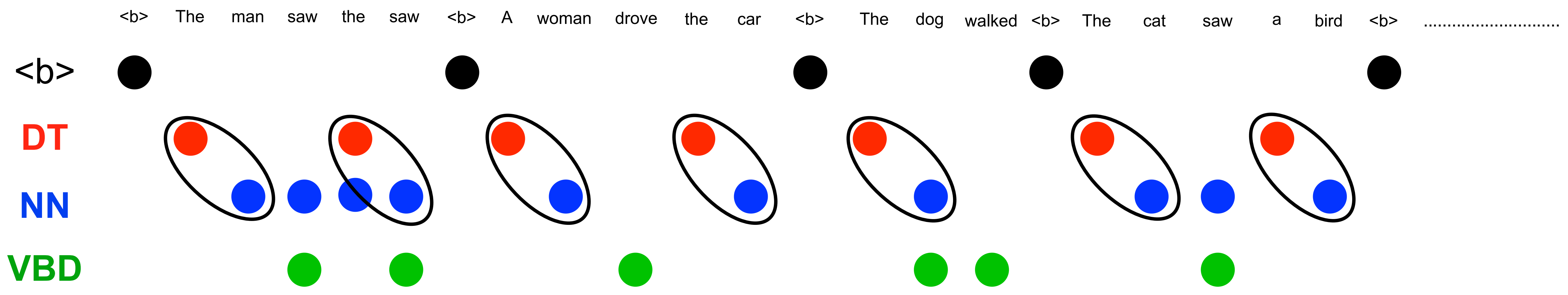


Model Minimization


<b> The man saw the saw <b> A woman drove the car <b> The dog walked <b> The cat saw a bird <b> ....
<b>
DT
NN
VBD

$f(\text{NN} \rightarrow \text{VBD}) \begin{cases} \text{tag bigram occurrences} \\ \text{weights on their nodes} \end{cases}$

Model Minimization



<b> The man saw the saw <b> A woman drove the car <b> The dog walked <b> The cat saw a bird <b> ....

<b>

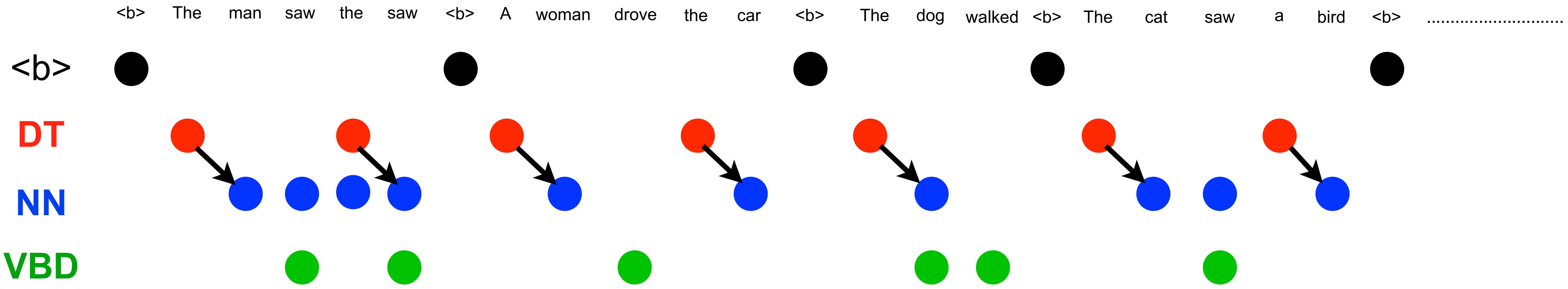
DT

NN

VBD

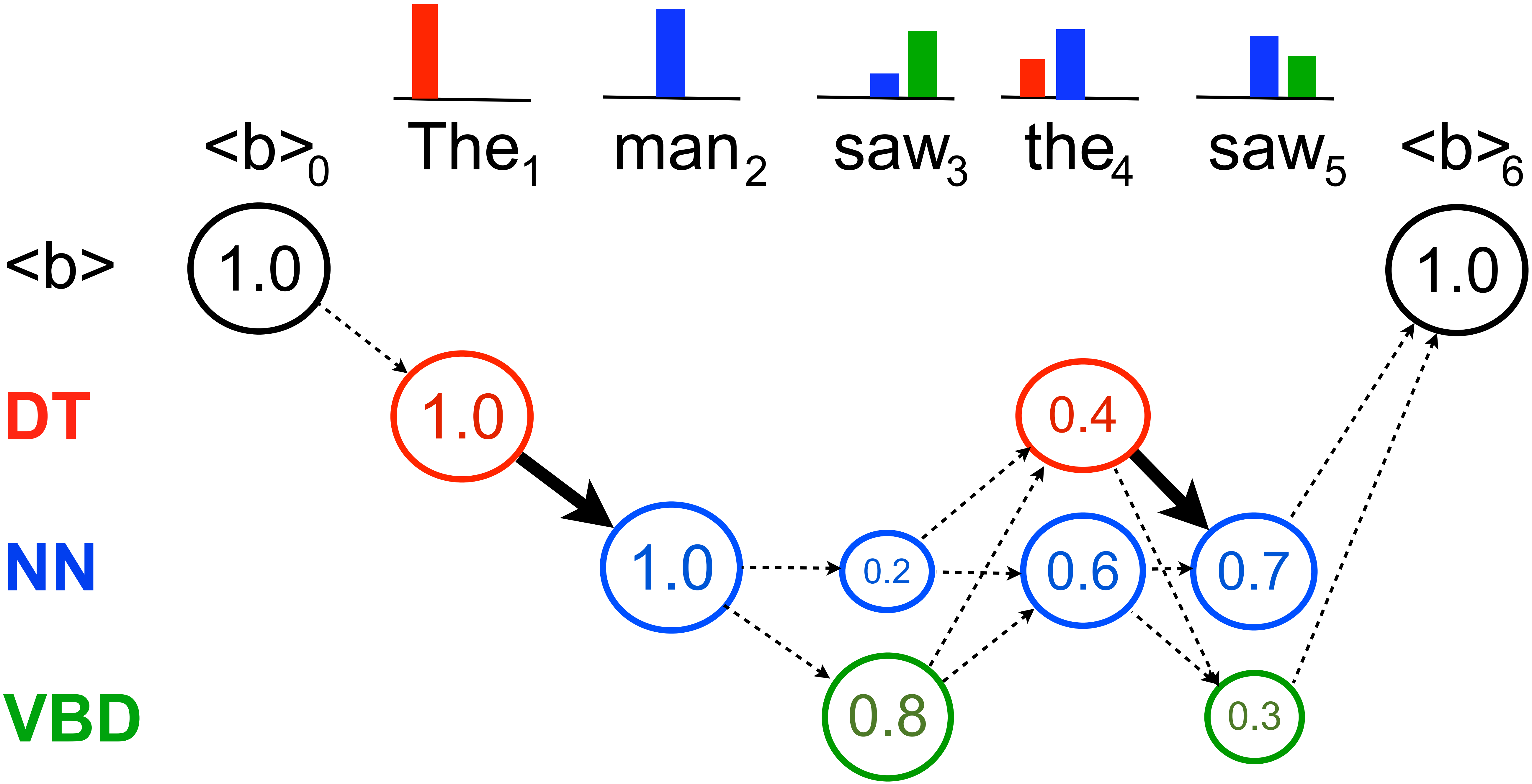
f(DT → NN) ✓

Model Minimization

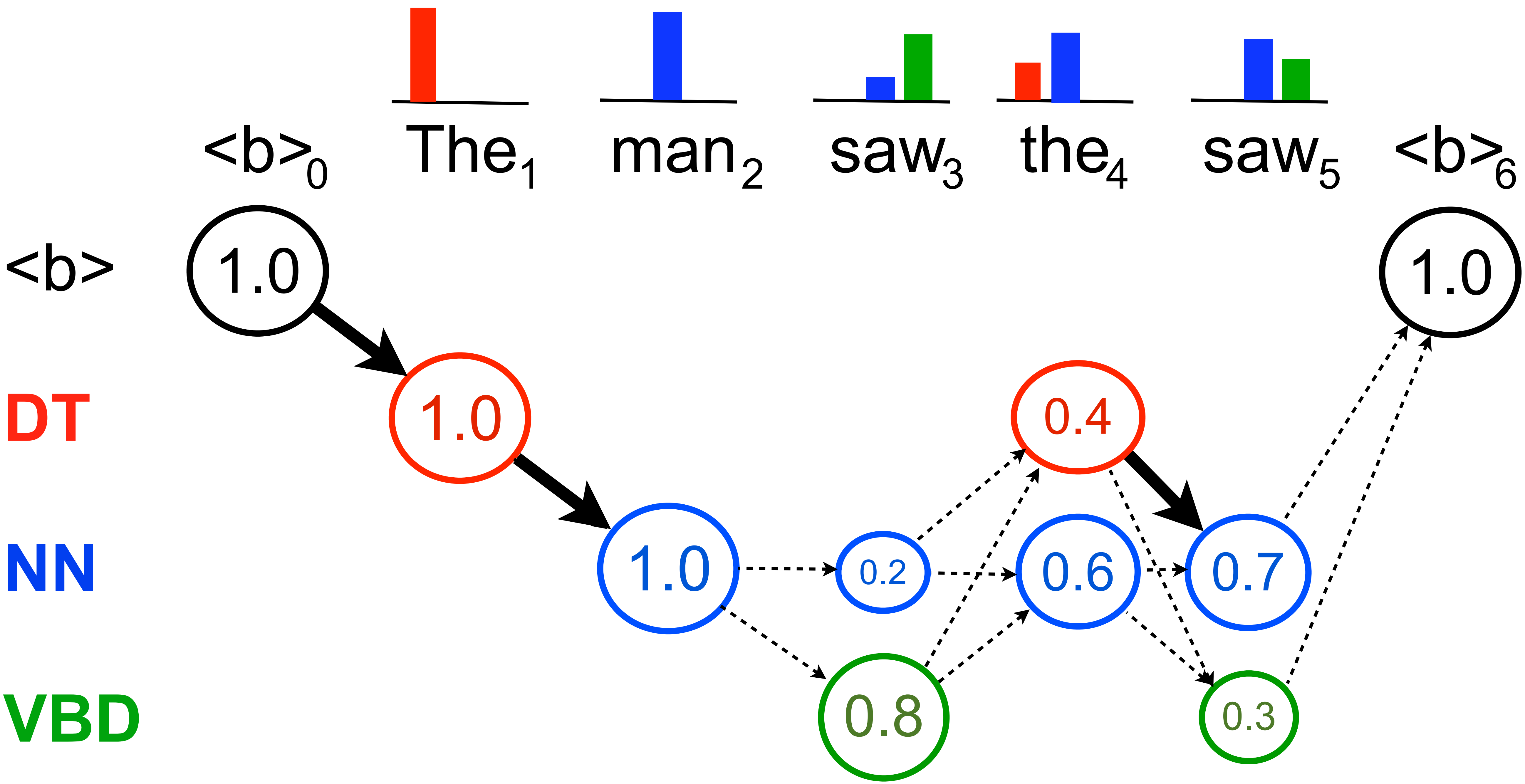


<b> The man saw the saw <b> A woman drove the car <b> The dog walked <b> The cat saw a bird <b> ....
<b>
DT
NN
VBD

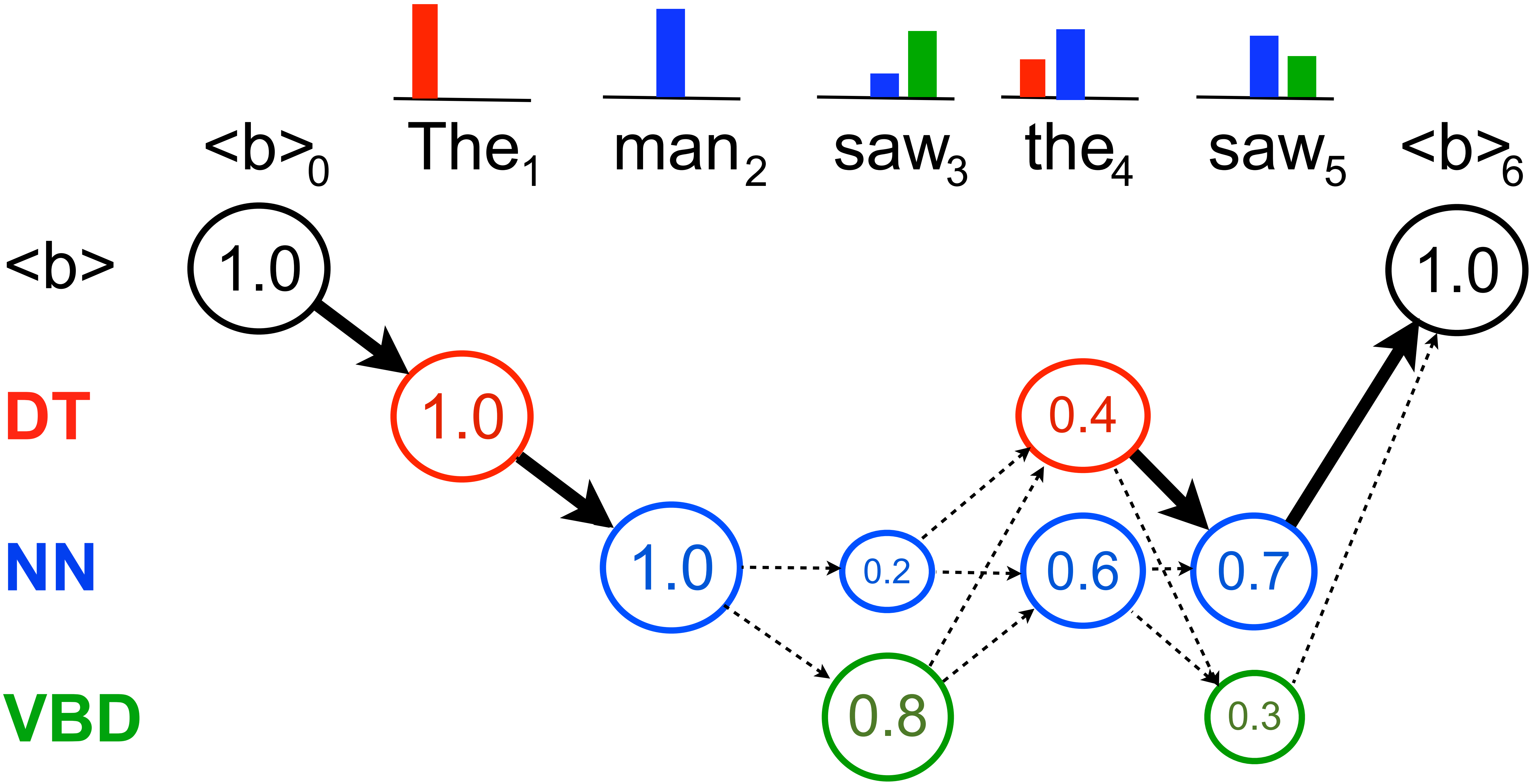
Model Minimization



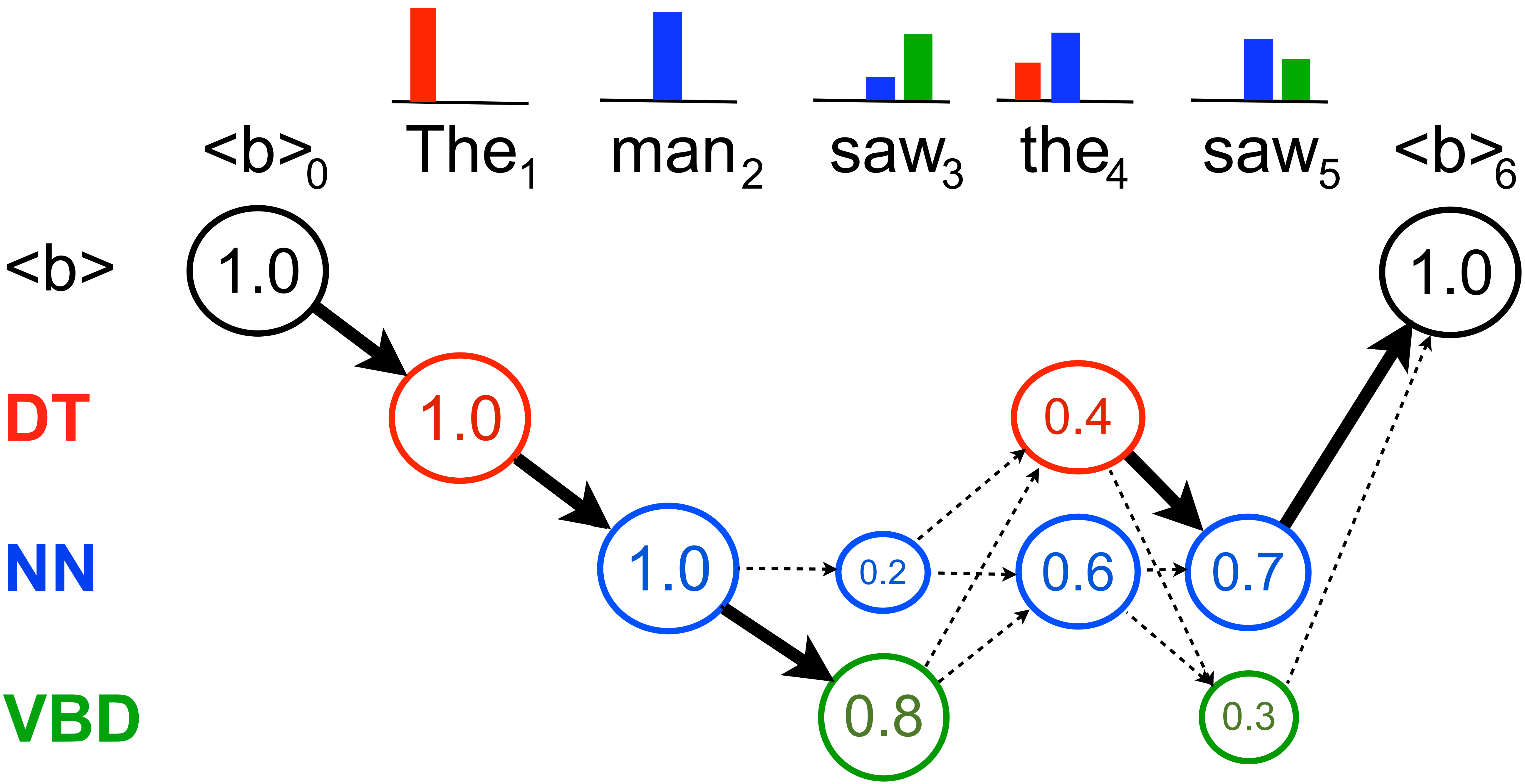
Model Minimization



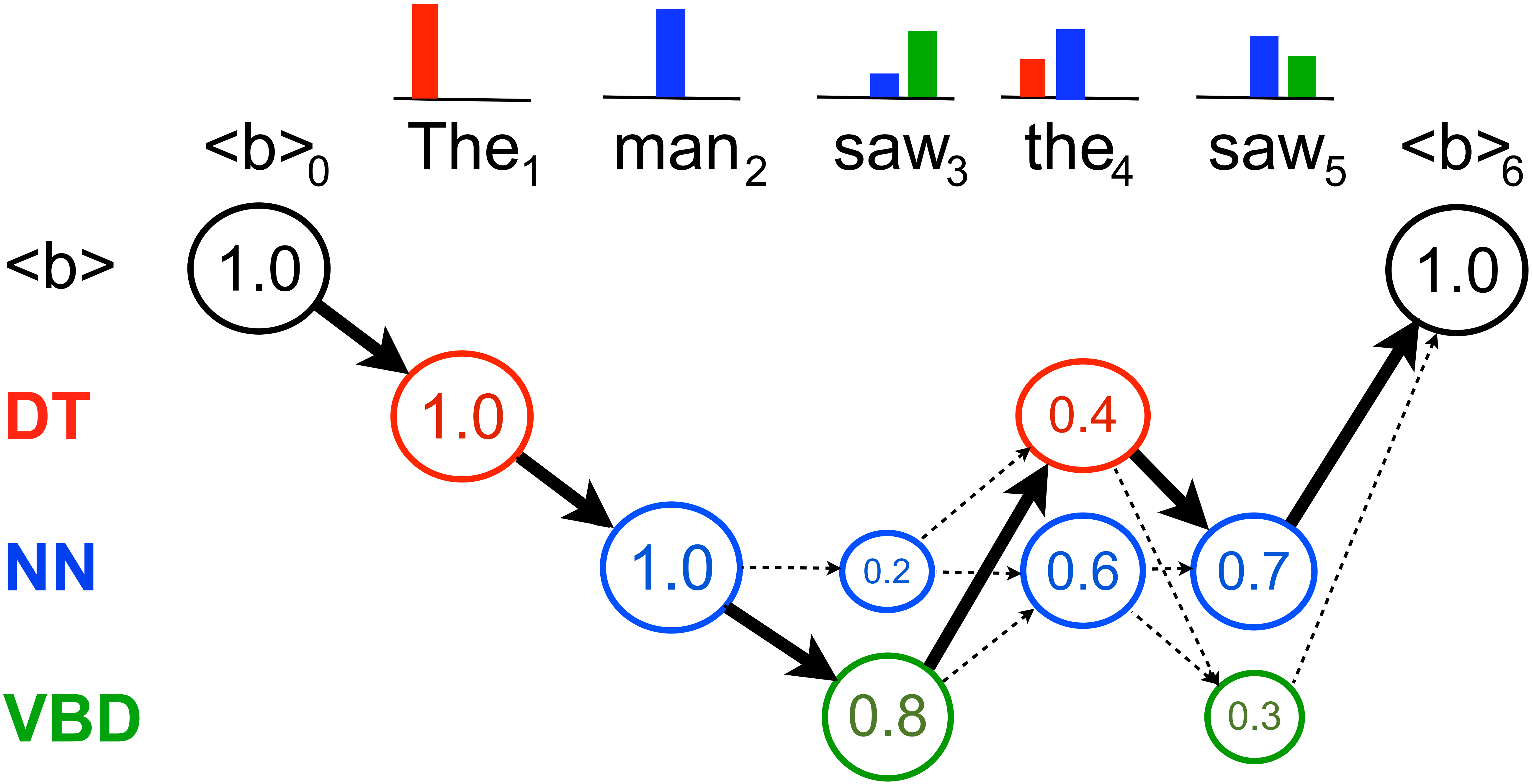
Model Minimization



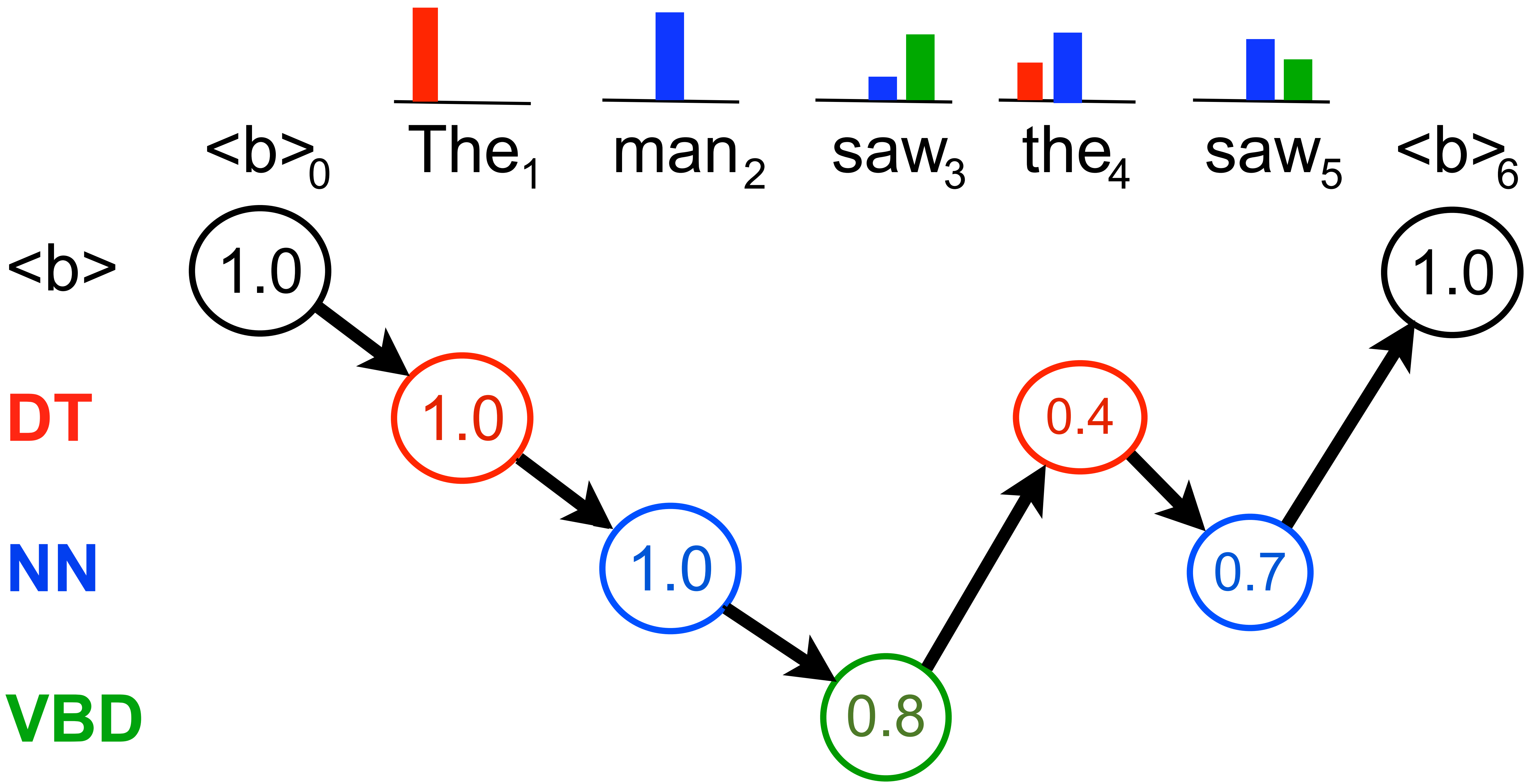
Model Minimization



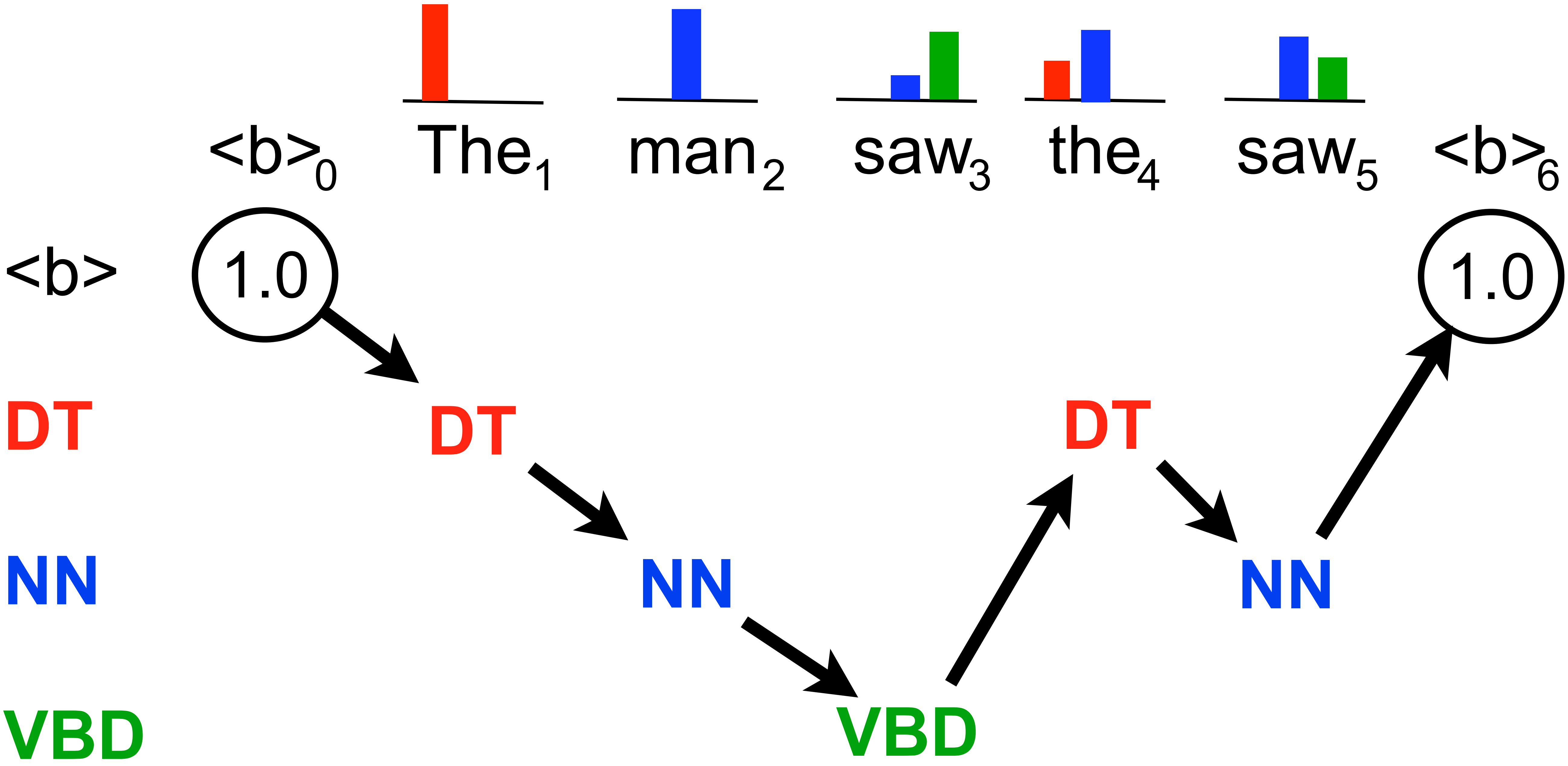
Model Minimization



Model Minimization



Model Minimization



<b>$_0$ The$_1$ man$_2$ saw$_3$ the$_4$ saw$_5$ <b>$_6$

<b>

DT

NN

VBD

1.0

DT

NN

VBD

DT

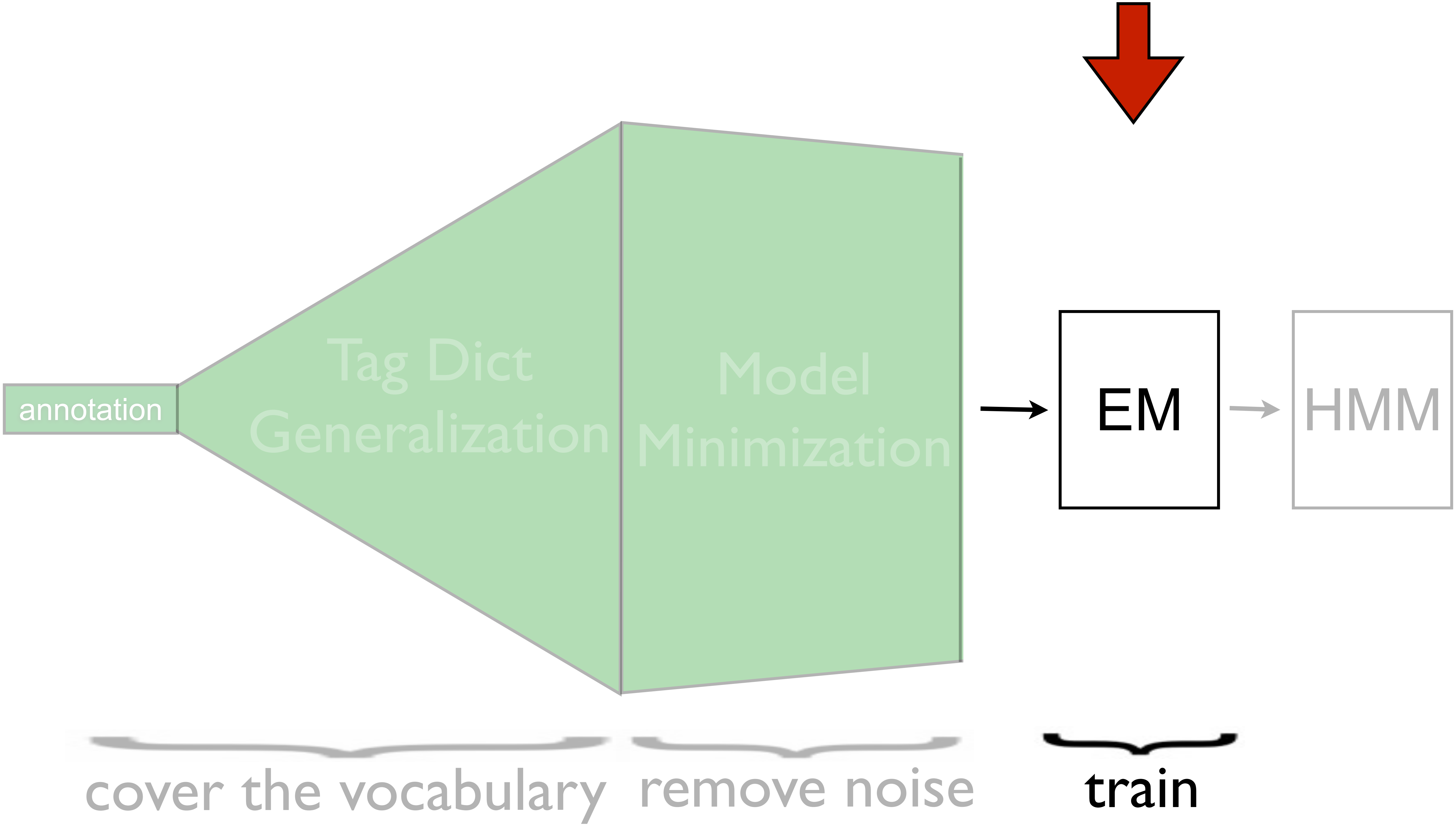
NN

1.0

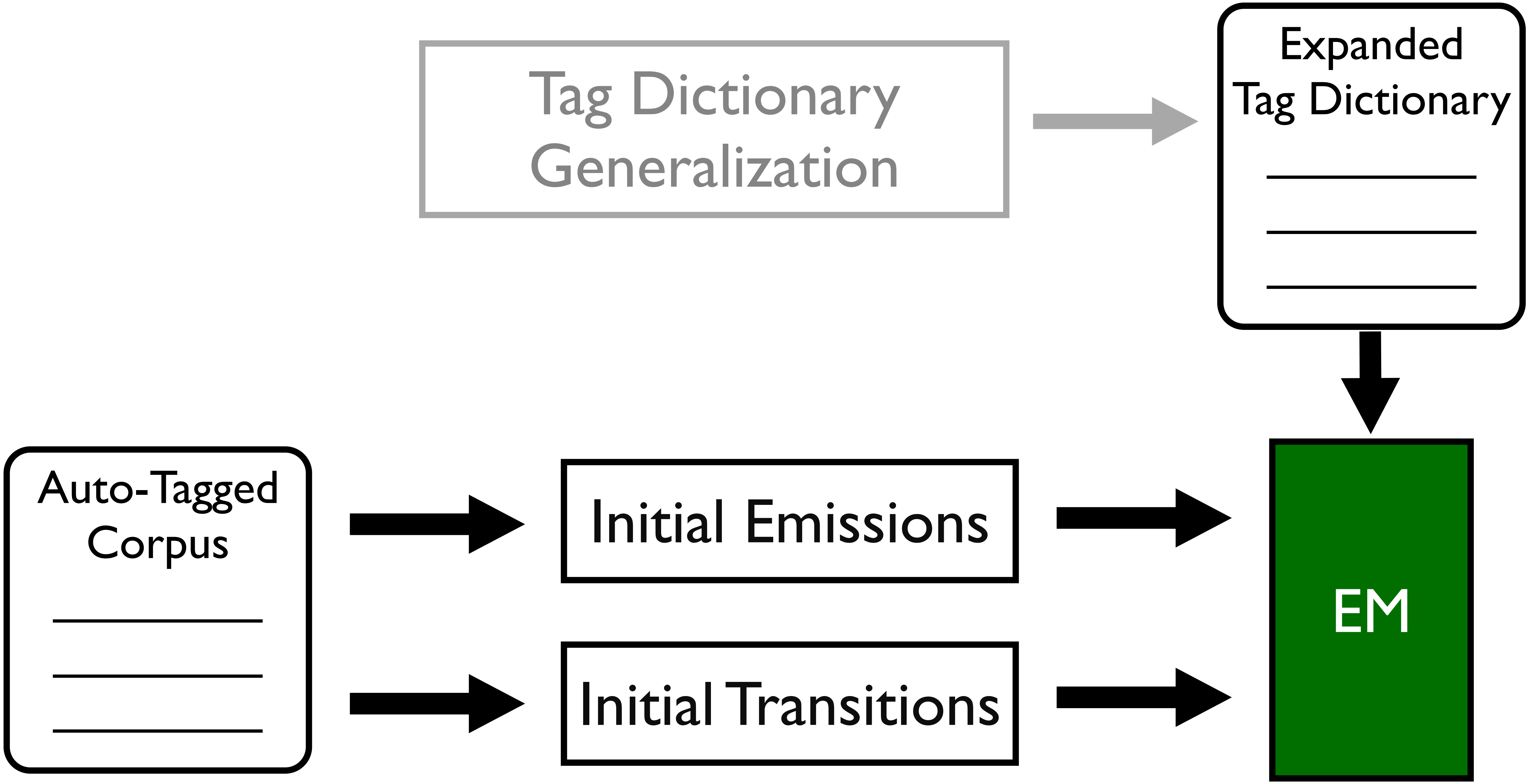
Model Minimization

<b>$_0$	The$_1$	man$_2$	saw$_3$	the$_4$	saw$_5$	<b>$_6$
	DT	**NN**	**VBD**	**DT**	**NN**	

Our Approach



Model Minimization



Tag Dictionary Generalization

Expanded Tag Dictionary

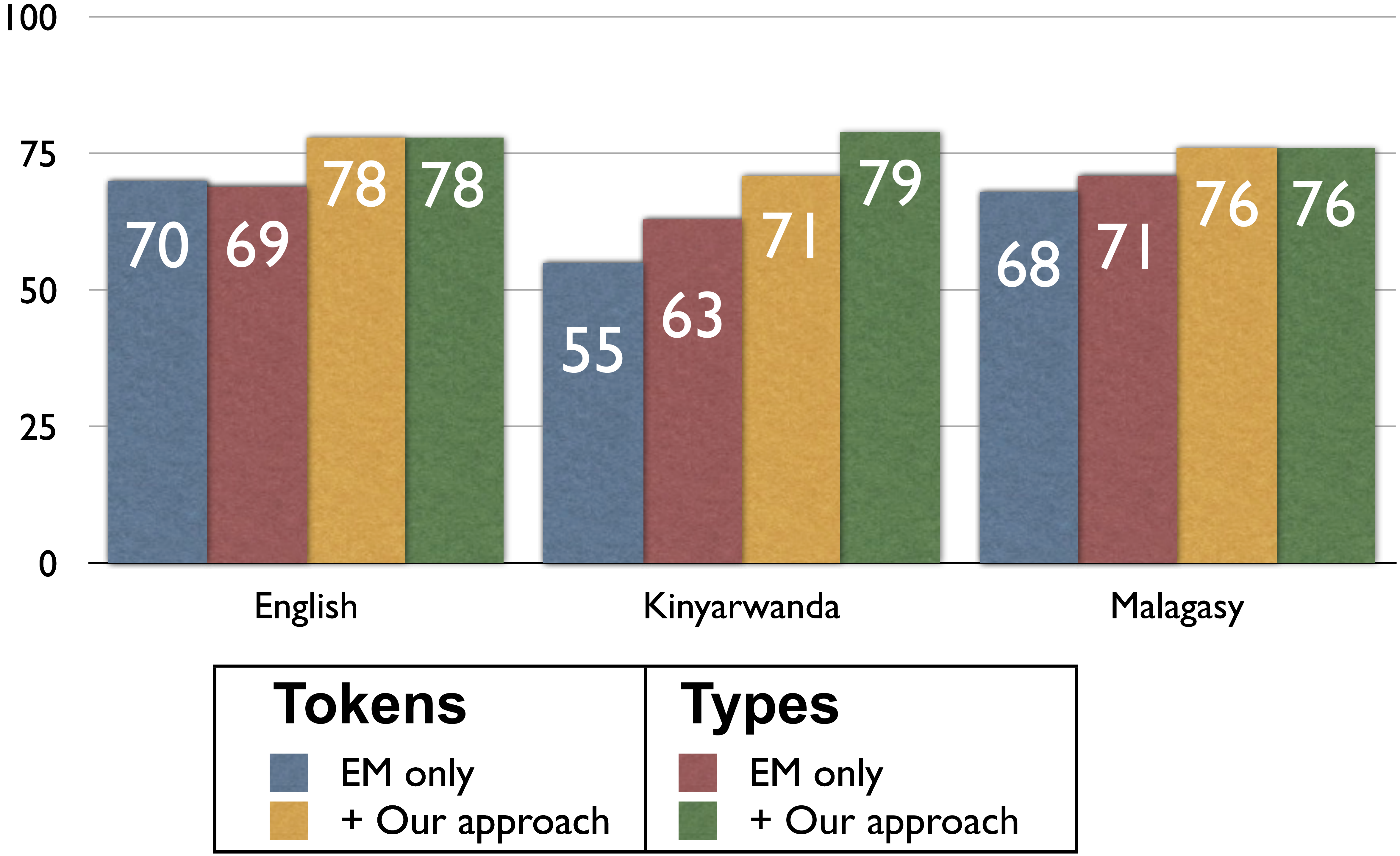
Auto-Tagged Corpus

Initial Emissions

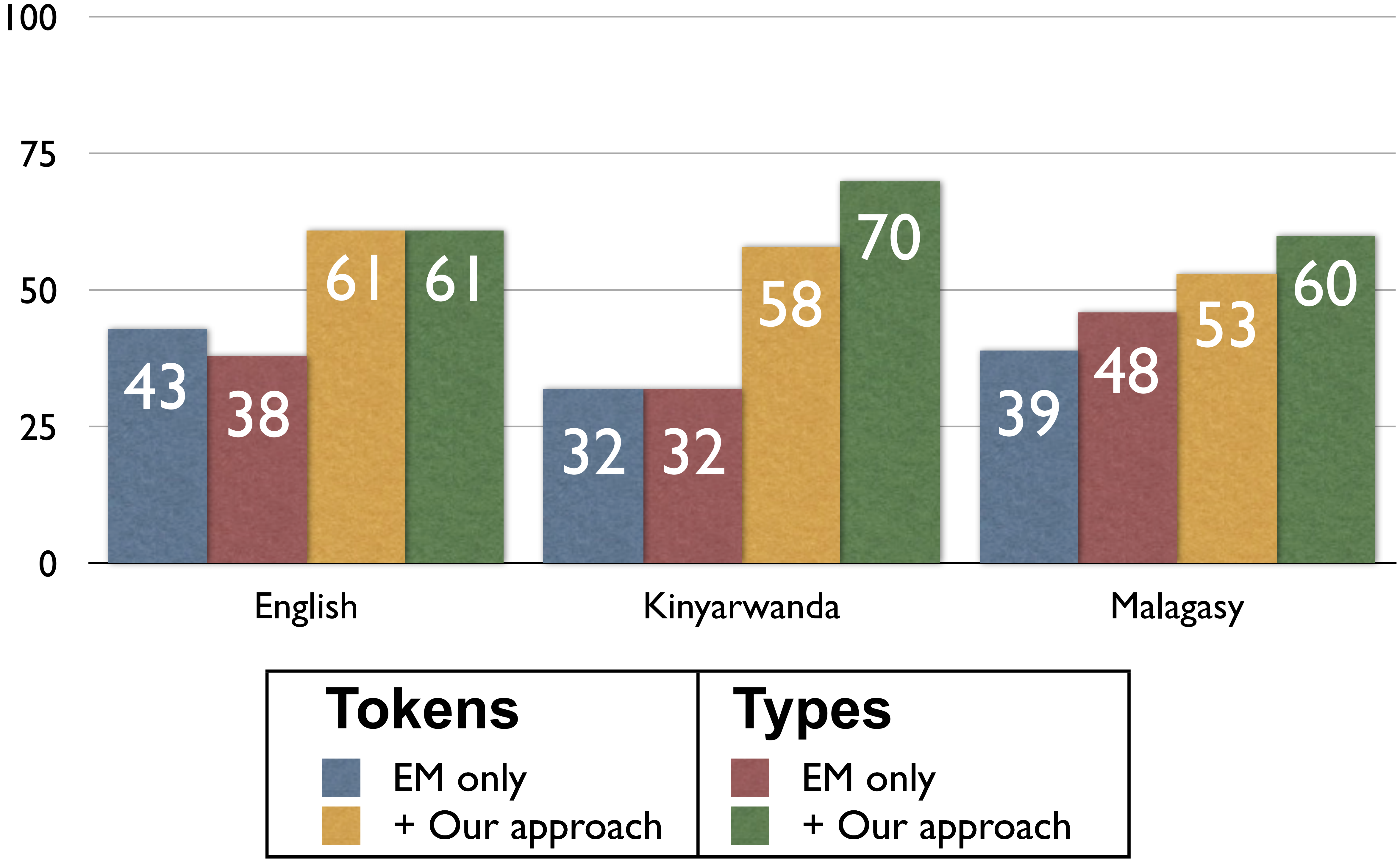
Initial Transitions

EM

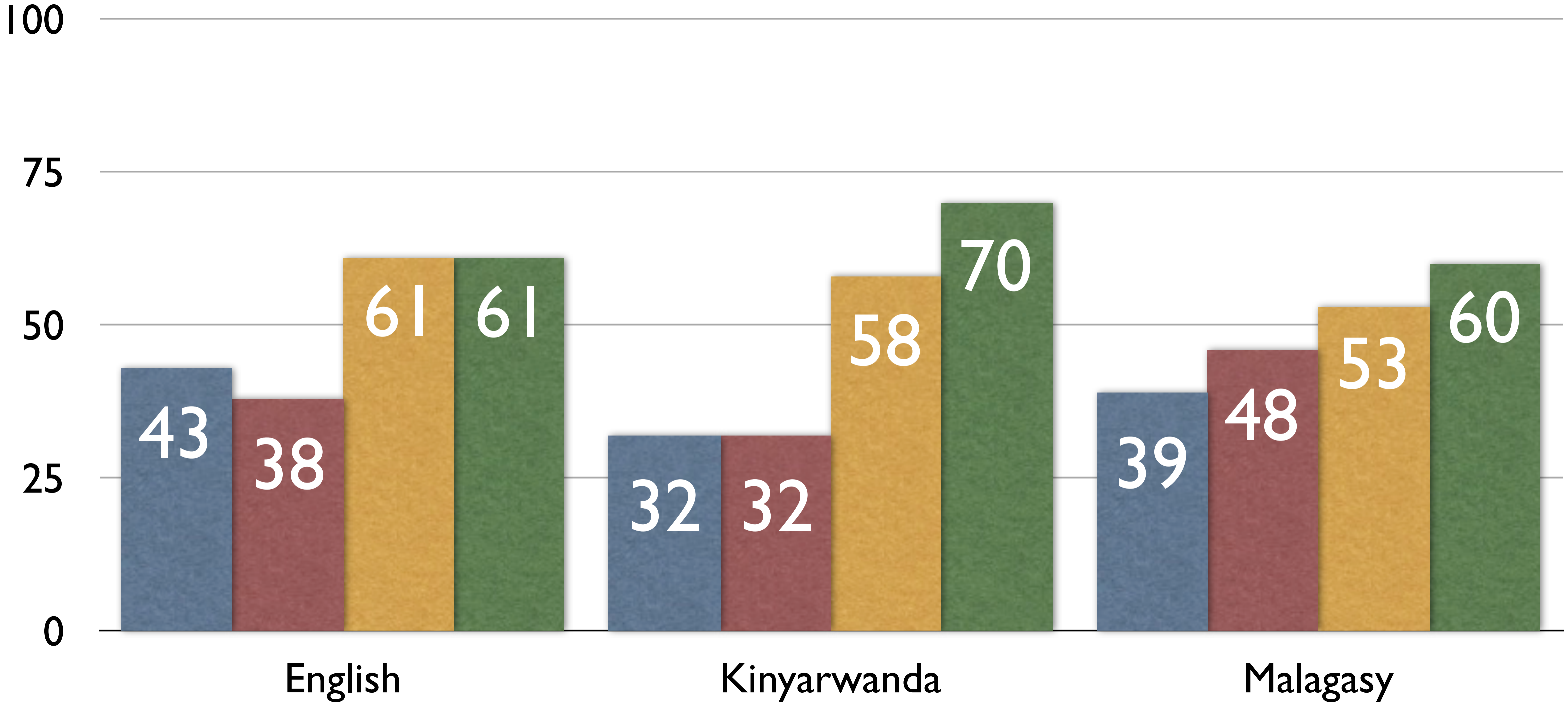
Total Accuracy



Unknown Accuracy



Unknown Accuracy



Remember: Very high unknown rates.

Especially for morphological-rich Kinyarwanda.

Conclusion

- Developed a semi-supervised approach to learn a tagger from realistically minimal input.
- Currently being used for further low-resource research (e.g. unsupervised dependency parsing).

ACL Preview

- Learning curves for annotation time
- Mixed types and tokens under fixed time constraints
- Morphological transducers
- **90%** accuracy on full 45 tag English Penn Treebank with **4 hours** of data

Software Available

Train your own low-resource taggers.

Or use our Kinyarwanda and Malagasy models.

Open source: link on my website or in the paper.