Product of a Grammar with an n-Gram Model for Statistical Machine Translation

Tobias Denkinger
tobias.denkinger@mailbox.tu-dresden.de

Lehrstuhl für Grundlagen der Programmierung
Institut für Theoretische Informatik
Technische Universität Dresden

June 7, 2013

Outline

Modelling and Training

translation function h: $F \rightarrow E$

Modelling and Training

translation function $h\colon F \to E$ (set of hypotheses $\mathcal{H} \subseteq E^F$)

$\mathcal{M} \subseteq \mathcal{M}(E \mid F)$

Modelling and Training

translation function $h\colon F \to E$ (set of hypotheses $\mathcal{H} \subseteq E^F$)

$$\mathcal{M} \subseteq \mathcal{M}(E \mid F)$$

$$\downarrow$$

$$\mathcal{H} = \{h_p \mid p \in \mathcal{M}\}$$

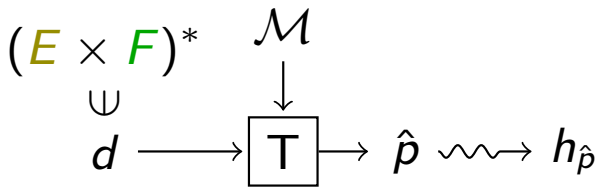
$$h_p(f) = \text{argmax}_{e \in E}\, p(e \mid f)$$

Modelling and Training

translation function $h\colon F \to E$ (set of hypotheses $\mathcal{H} \subseteq E^F$)

$$\mathcal{M} \subseteq \mathcal{M}(E \mid F)$$

$$\mathcal{H} = \{h_p \mid p \in \mathcal{M}\}$$
$$h_p(f) = \operatorname{argmax}_{e \in E} p(e \mid f)$$



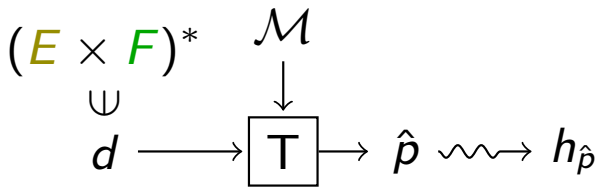
Modelling and Training

translation function $h\colon F \to E$ (set of hypotheses $\mathcal{H} \subseteq E^F$)

$$\mathcal{M} \subseteq \mathcal{M}(E \mid F)$$

$$\mathcal{H} = \{h_p \mid p \in \mathcal{M}\}$$
$$h_p(f) = \operatorname{argmax}_{e \in E} p(e \mid f)$$

$$\mathcal{H} = \{h_{p_1,p_2} \mid p_1 \in \mathcal{M}_1, p_2 \in \mathcal{M}_2\}$$
$$h_{p_1,p_2}(f) = \operatorname{argmax}_{e \in E} p_1(f \mid e) \cdot p_2(e)$$

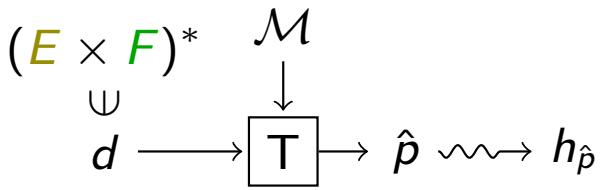


Modelling and Training

translation function $h\colon F \to E$ (set of hypotheses $\mathcal{H} \subseteq E^F$)

$$\mathcal{M} \subseteq \mathcal{M}(E \mid F)$$
$$\downarrow$$
$$\mathcal{H} = \{h_p \mid p \in \mathcal{M}\}$$
$$h_p(f) = \operatorname{argmax}_{e \in E} p(e \mid f)$$

$$\mathcal{M}_1 \subseteq \mathcal{M}(F \mid E)$$
$$\mathcal{M}_2 \subseteq \mathcal{M}(E)$$
$$\downarrow$$
$$\mathcal{H} = \{h_{p_1,p_2} \mid p_1 \in \mathcal{M}_1, p_2 \in \mathcal{M}_2\}$$
$$h_{p_1,p_2}(f) = \operatorname{argmax}_{e \in E} p_1(f \mid e) \cdot p_2(e)$$



Modelling and Training

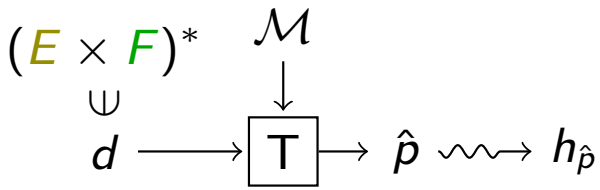
translation function $h\colon F \to E$ (set of hypotheses $\mathcal{H} \subseteq E^F$)

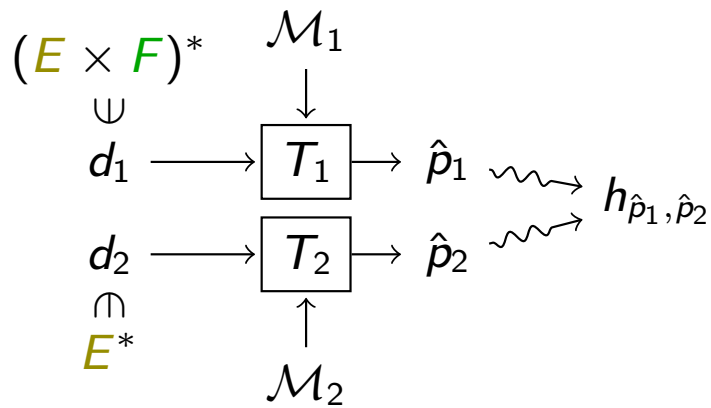
$$\mathcal{M} \subseteq \mathcal{M}(E \mid F)$$

$$\mathcal{H} = \{h_p \mid p \in \mathcal{M}\}$$
$$h_p(f) = \operatorname{argmax}_{e \in E} p(e \mid f)$$

$$\mathcal{M}_1 \subseteq \mathcal{M}(F \mid E)$$
$$\mathcal{M}_2 \subseteq \mathcal{M}(E)$$

$$\mathcal{H} = \{h_{p_1,p_2} \mid p_1 \in \mathcal{M}_1, p_2 \in \mathcal{M}_2\}$$
$$h_{p_1,p_2}(f) = \operatorname{argmax}_{e \in E} p_1(f \mid e) \cdot p_2(e)$$



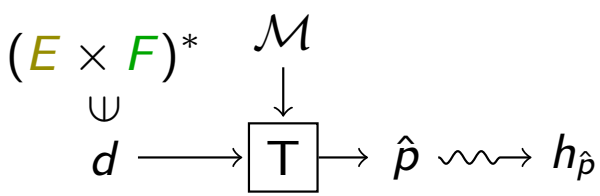


Modelling and Training

translation function $h\colon F \to E$ (set of hypotheses $\mathcal{H} \subseteq E^F$)

$$\mathcal{M} \subseteq \mathcal{M}(E \mid F)$$

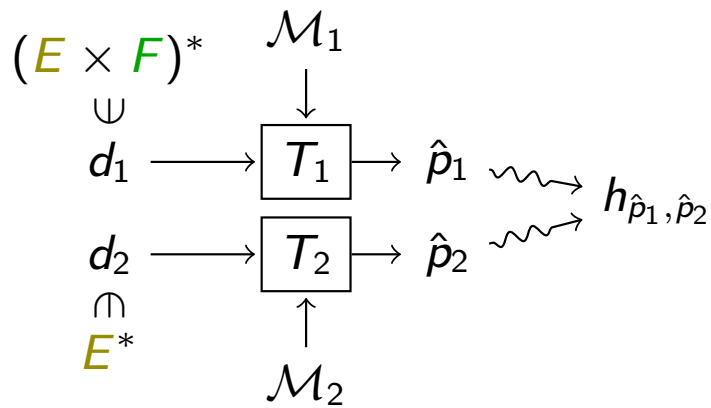
$$\mathcal{H} = \{h_p \mid p \in \mathcal{M}\}$$
$$h_p(f) = \operatorname{argmax}_{e \in E} p(e \mid f)$$



source-channel approach

$$\mathcal{M}_1 \subseteq \mathcal{M}(F \mid E)$$
$$\mathcal{M}_2 \subseteq \mathcal{M}(E)$$

$$\mathcal{H} = \{h_{p_1,p_2} \mid p_1 \in \mathcal{M}_1, p_2 \in \mathcal{M}_2\}$$
$$h_{p_1,p_2}(f) = \operatorname{argmax}_{e \in E} p_1(f \mid e) \cdot p_2(e)$$



Outline

Decoding

source-channel approach

$$h_{p_1,p_2}(f) = \text{argmax}_{e \in E} \underbrace{p_1}_{\in \mathcal{M}_1}(f \mid e) \cdot \underbrace{p_2}_{\in \mathcal{M}_2}(e)$$

Decoding

source-channel approach

$$h_{p_1,p_2}(f) = \text{argmax}_{e \in E} \underbrace{p_1}_{\in \mathcal{M}_1}(f \mid e) \cdot \underbrace{p_2}_{\in \mathcal{M}_2}(e)$$

- $\mathcal{M}_1$ . . . interpreted regular tree grammars (IRTGs)

Decoding

source-channel approach

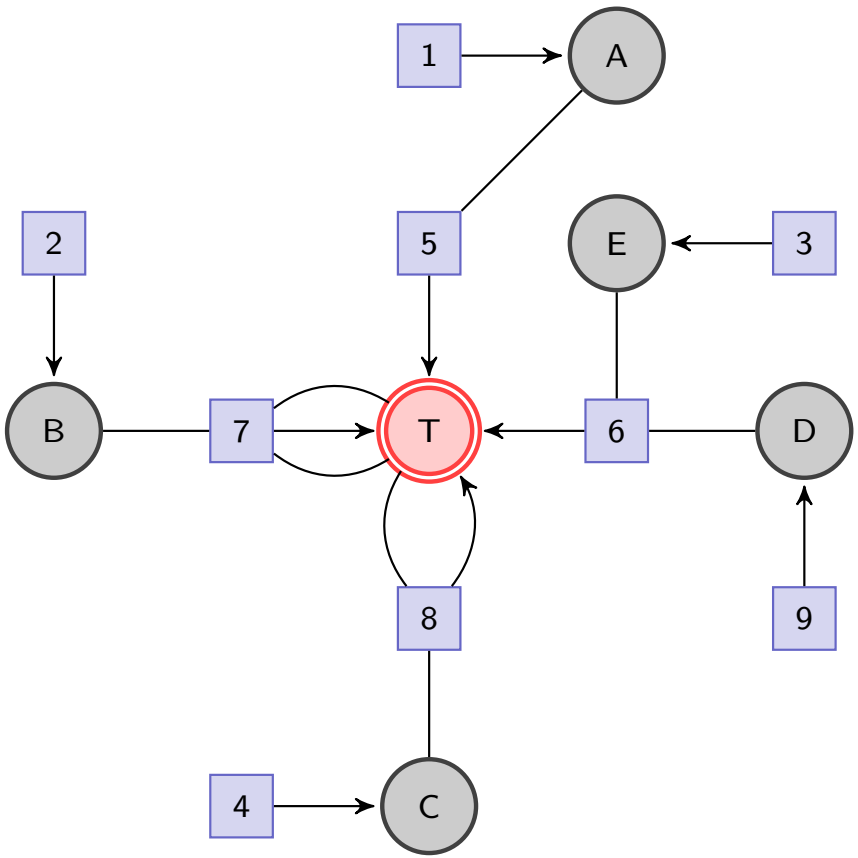
$$h_{p_1,p_2}(f) = \text{argmax}_{e \in E} \underbrace{p_1}_{\in \mathcal{M}_1}(f \mid e) \cdot \underbrace{p_2}_{\in \mathcal{M}_2}(e)$$

- $\mathcal{M}_1$ . . . interpreted regular tree grammars (IRTGs)
- $\mathcal{M}_2$ . . . n-gram model

Outline

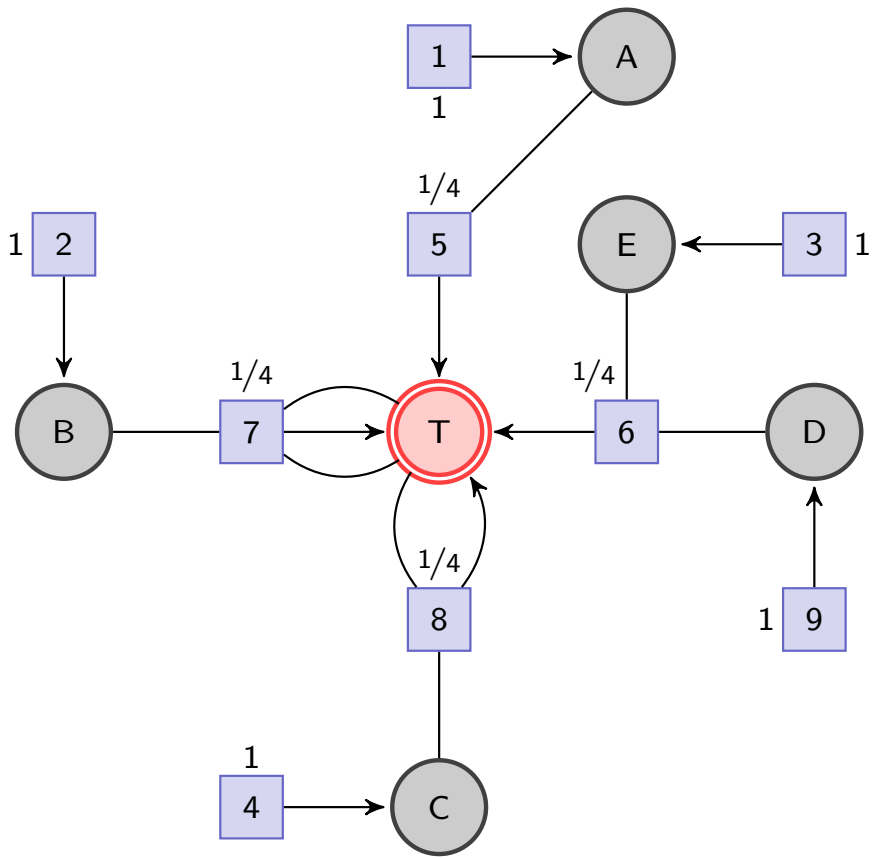
IRTG

- RTG (shown as hypergraph)



IRTG

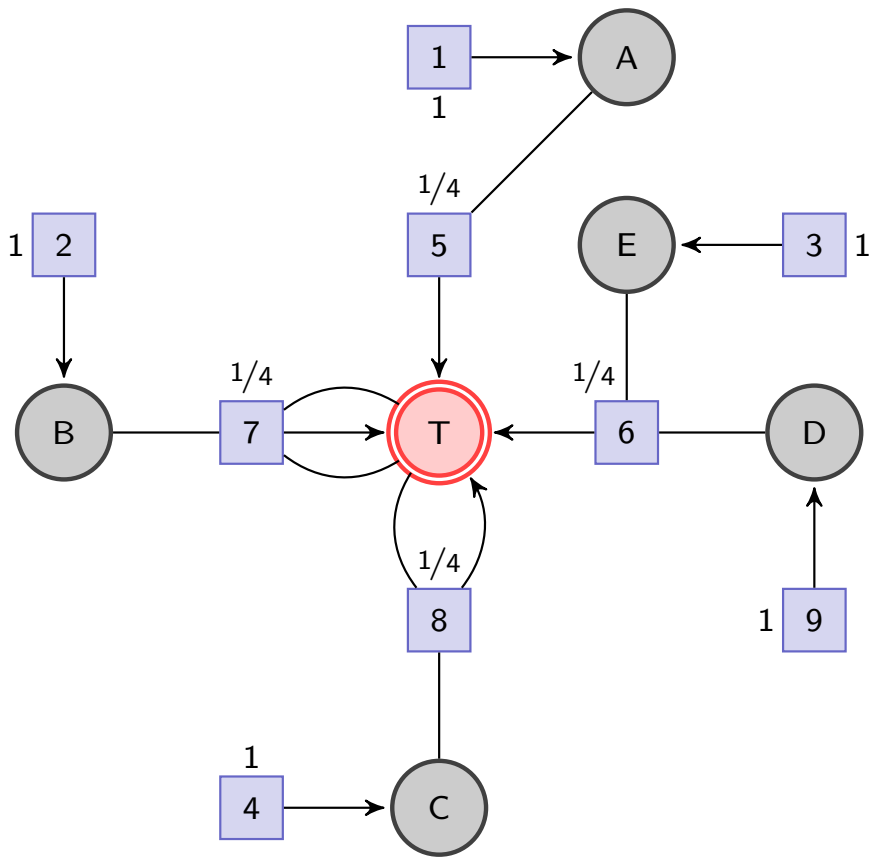
- wRTG (shown as hypergraph)



IRTG

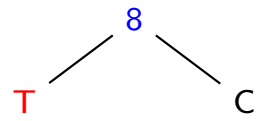
- wRTG (shown as hypergraph)

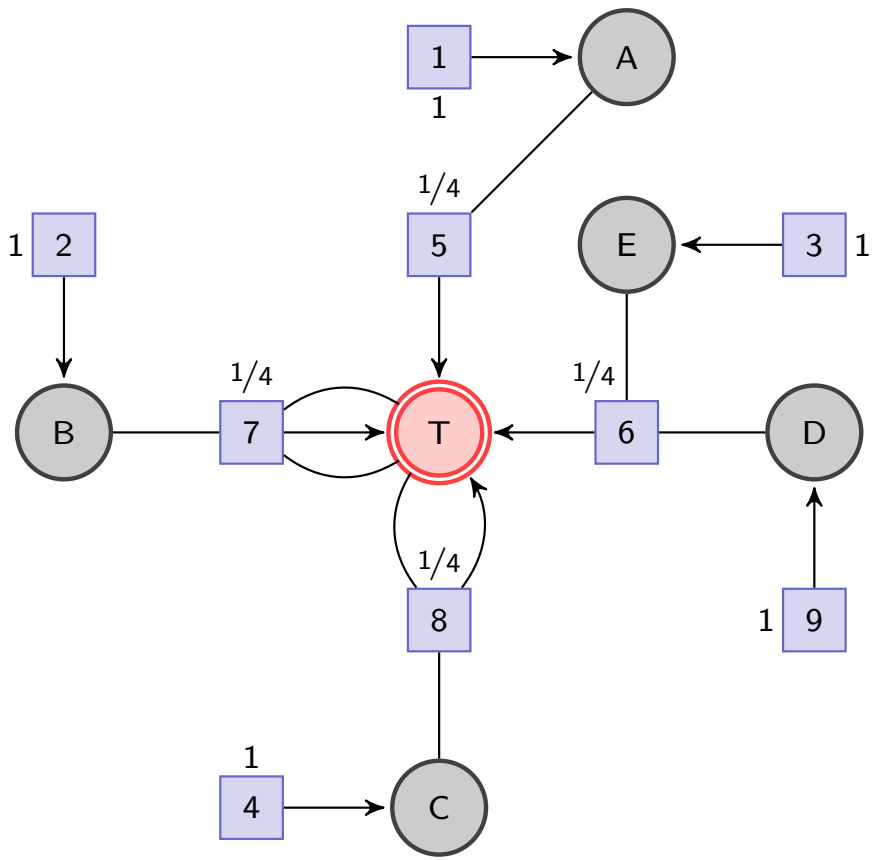
T



IRTG

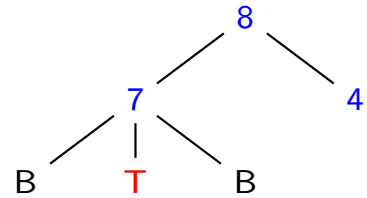
- wRTG (shown as hypergraph)

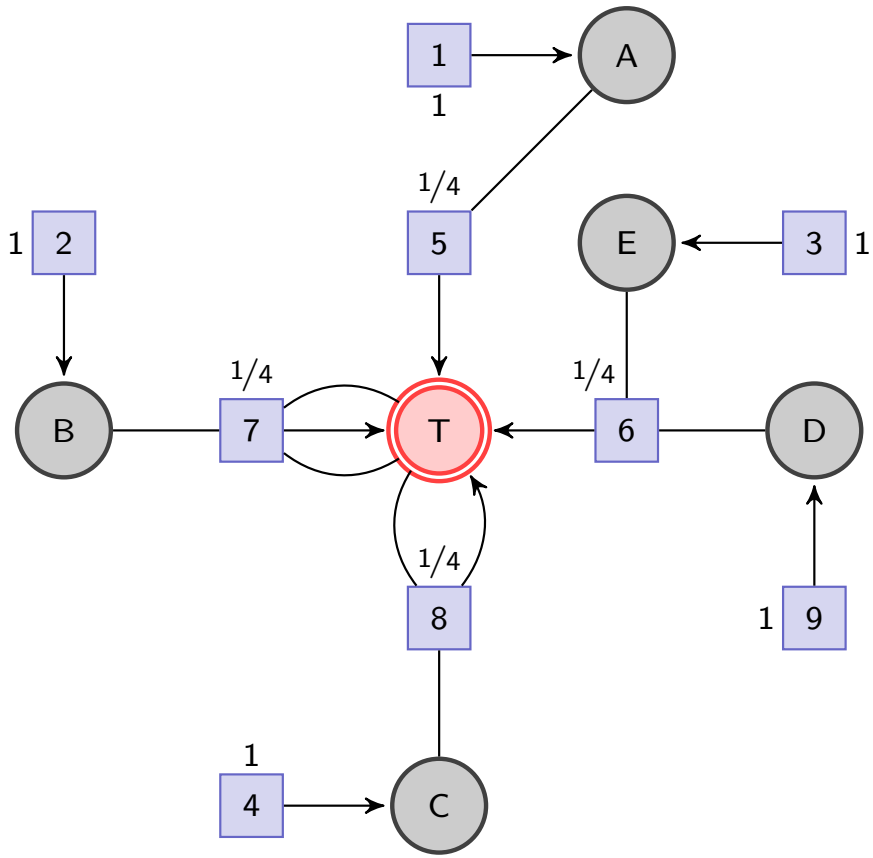




IRTG

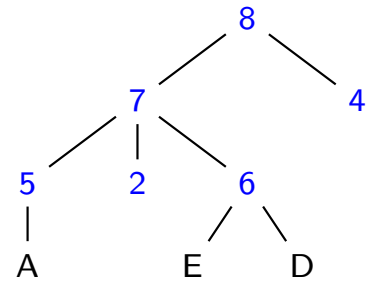
- wRTG (shown as hypergraph)

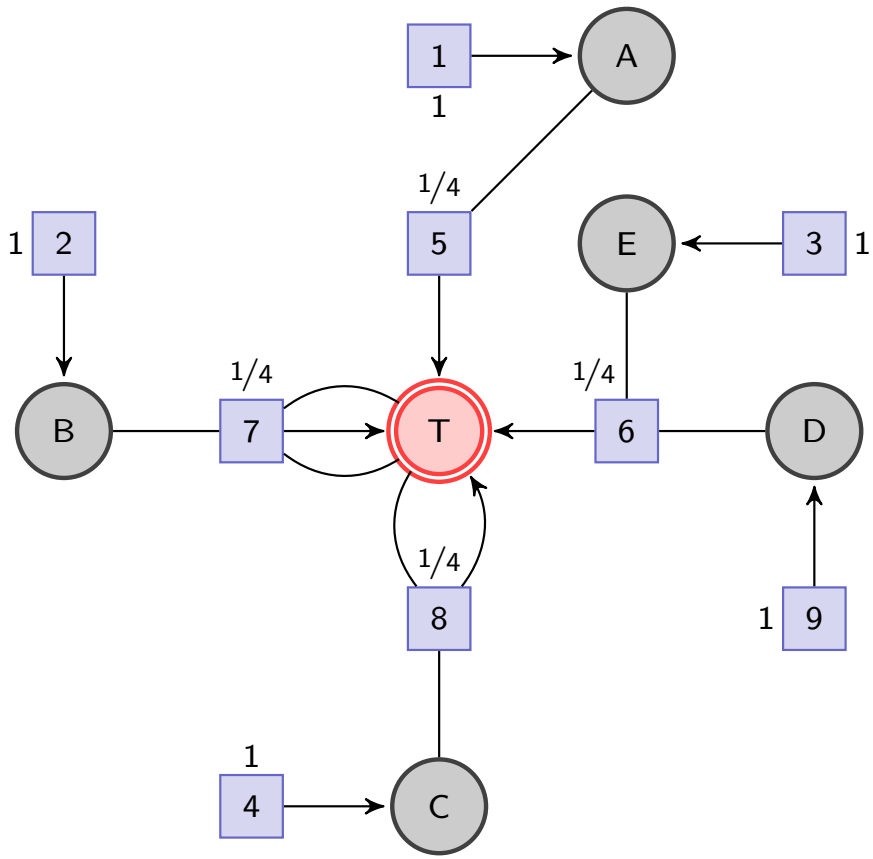




IRTG

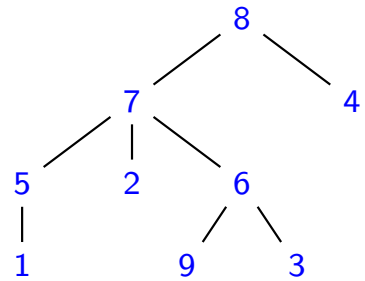
- wRTG (shown as hypergraph)

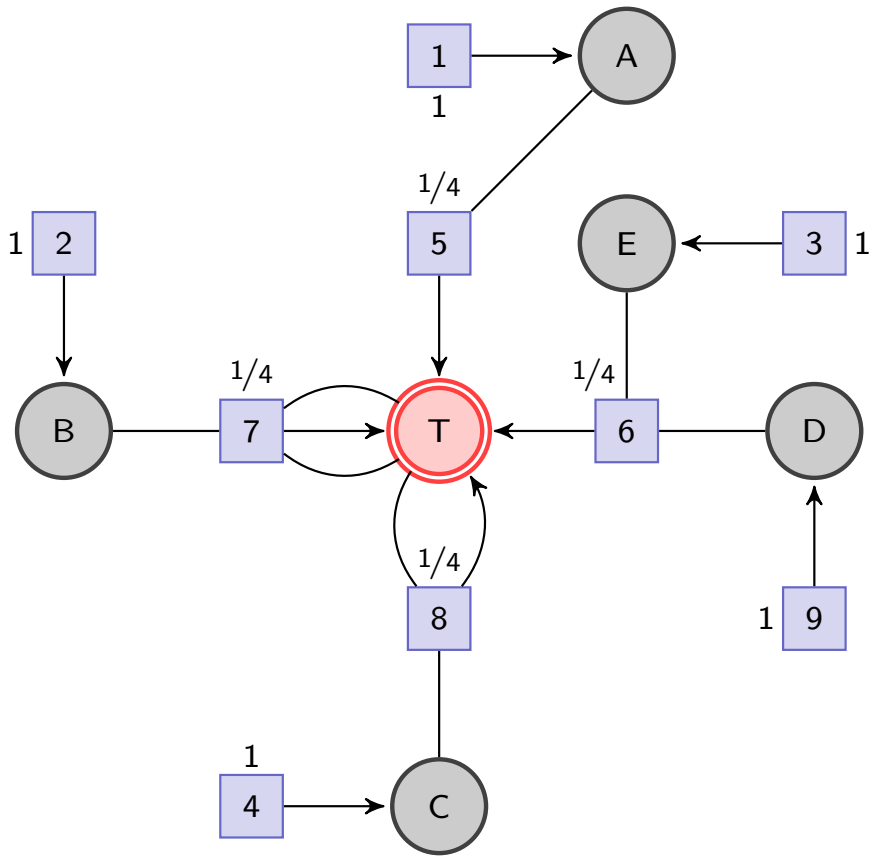




IRTG

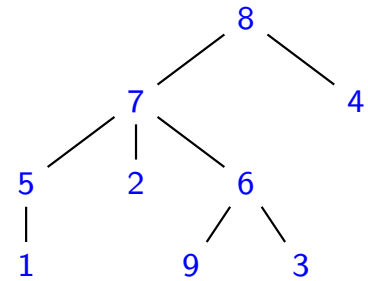
- wRTG (shown as hypergraph)

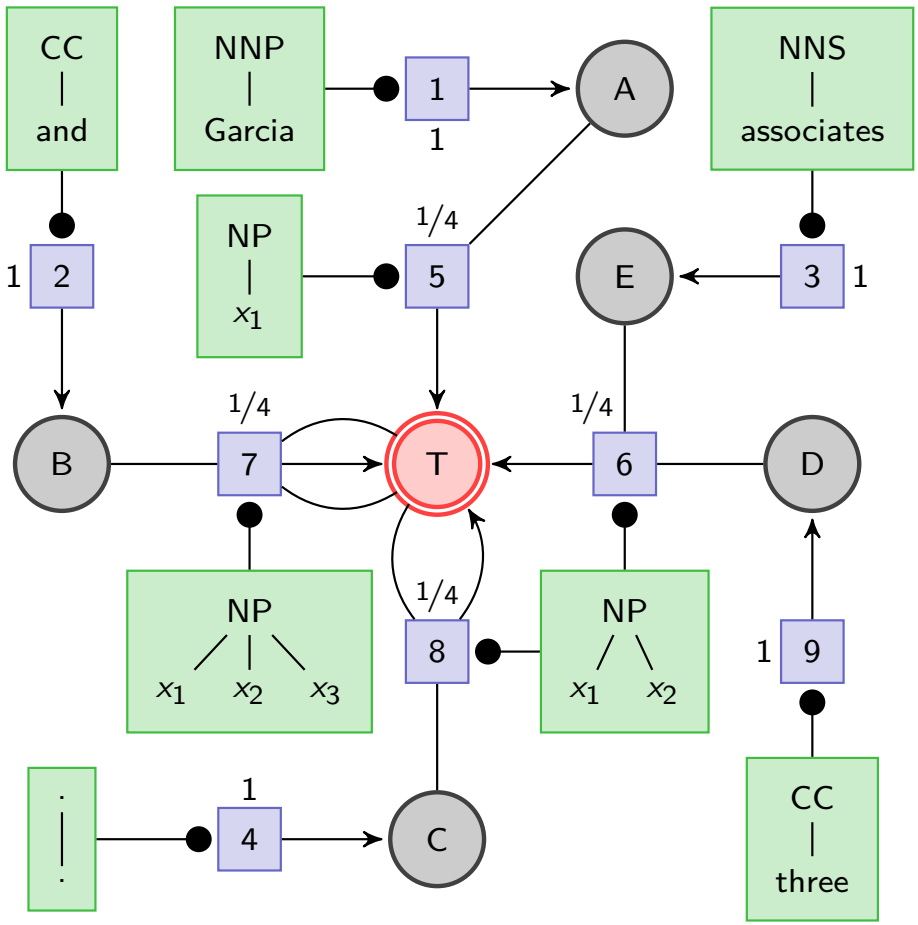




IRTG

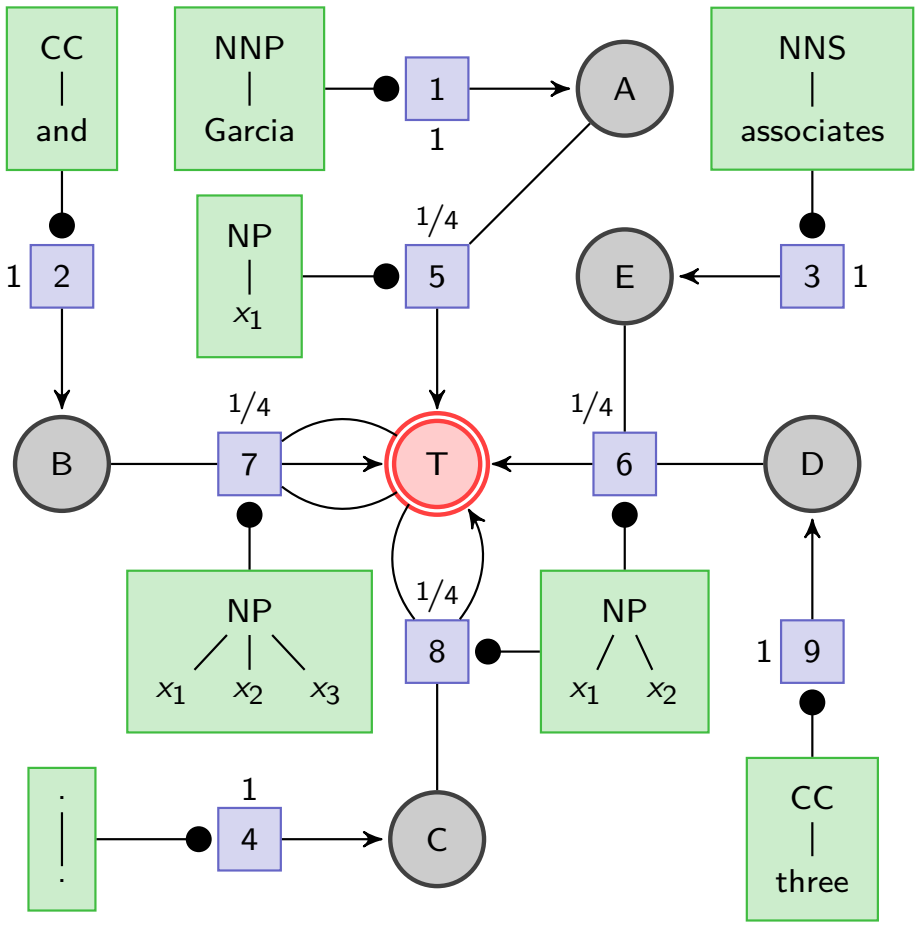
- wRTG (shown as hypergraph)
- tree homomorphism h_1

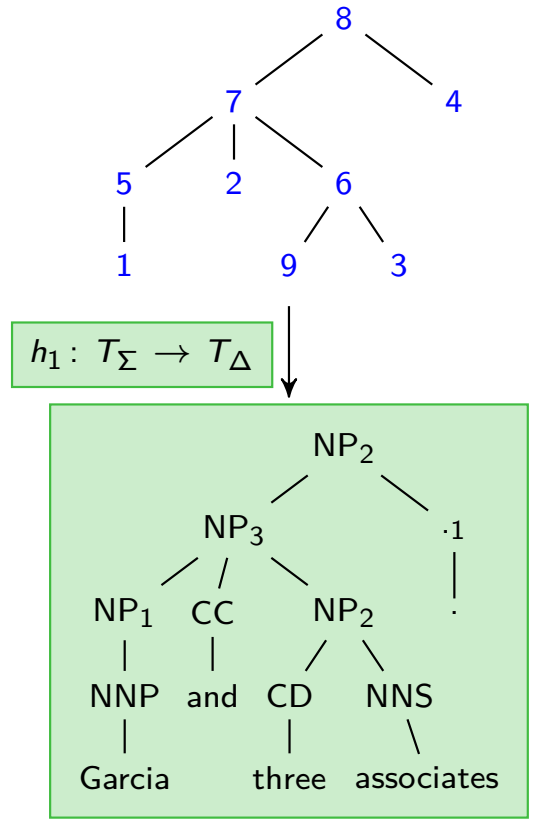




IRTG

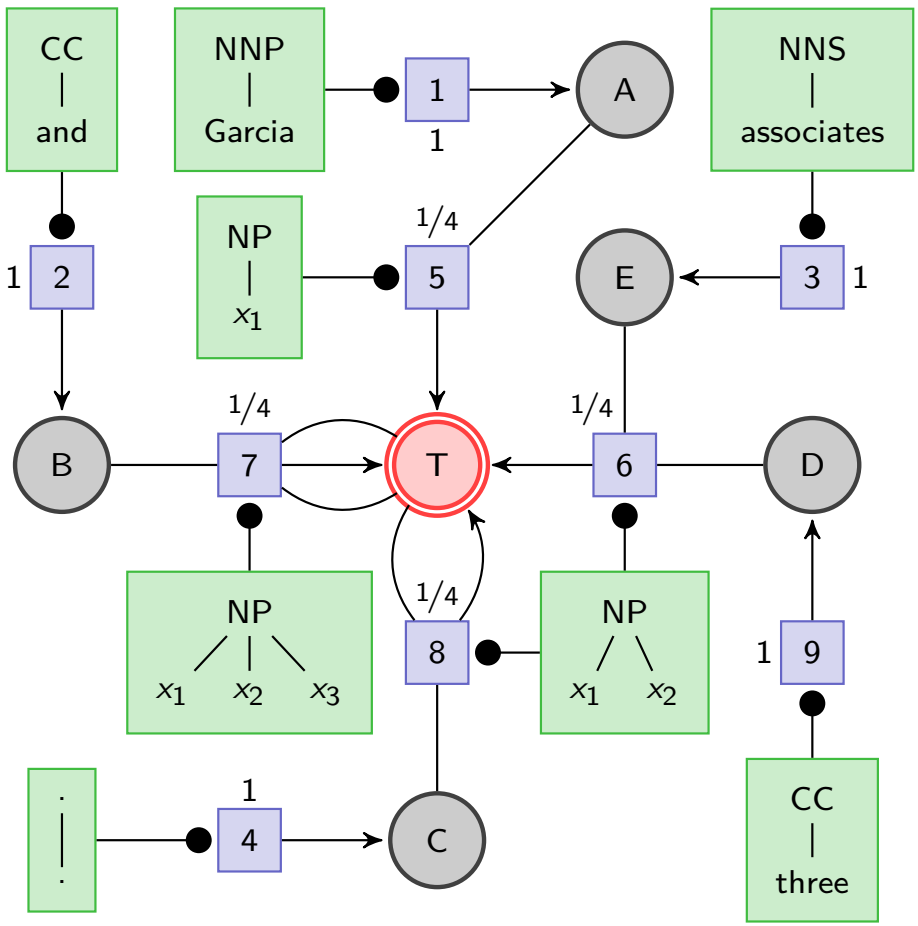
- wRTG (shown as hypergraph)
- tree homomorphism h_1

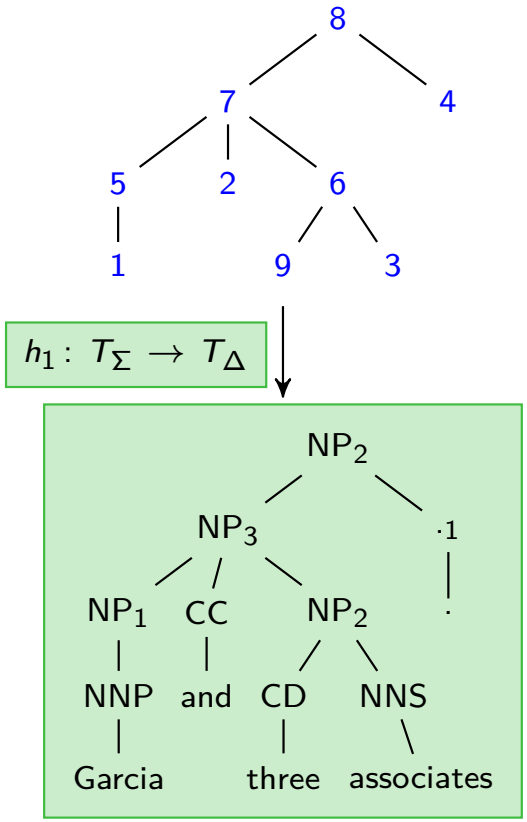




IRTG

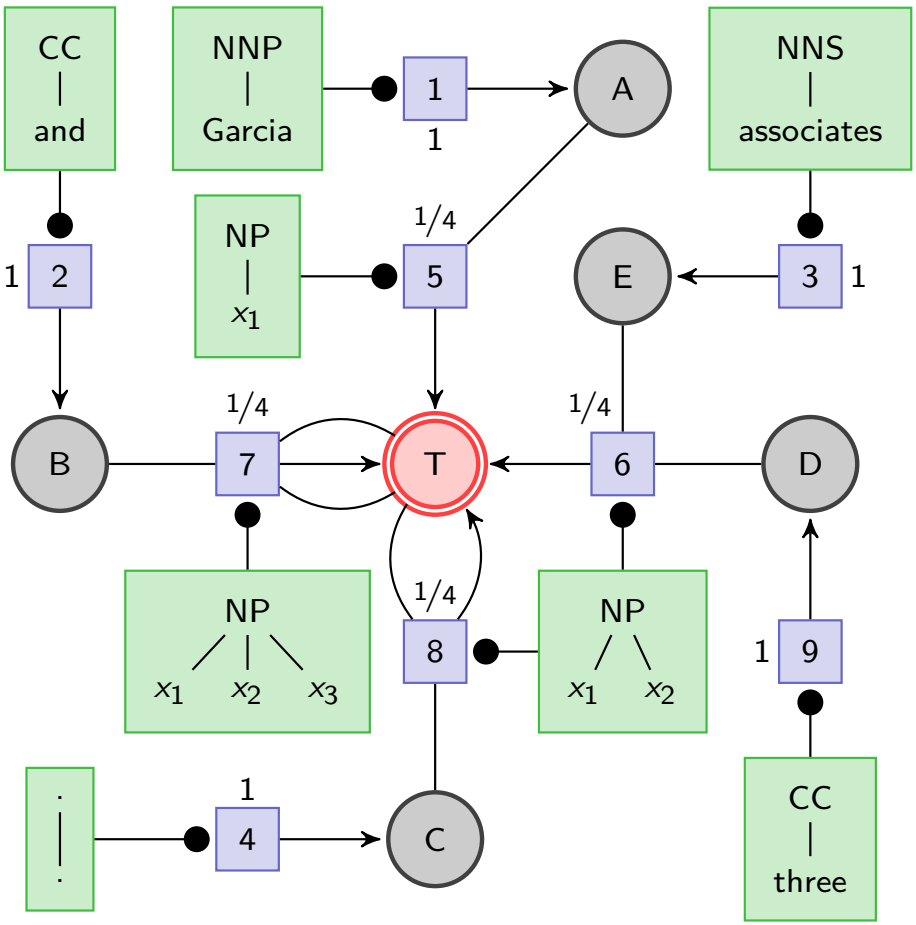
- wRTG (shown as hypergraph)
- tree homomorphism h_1
- algebra $\mathcal{A}_1 = \langle T_\Delta, (\cdot)^{\mathcal{A}_1} \rangle$

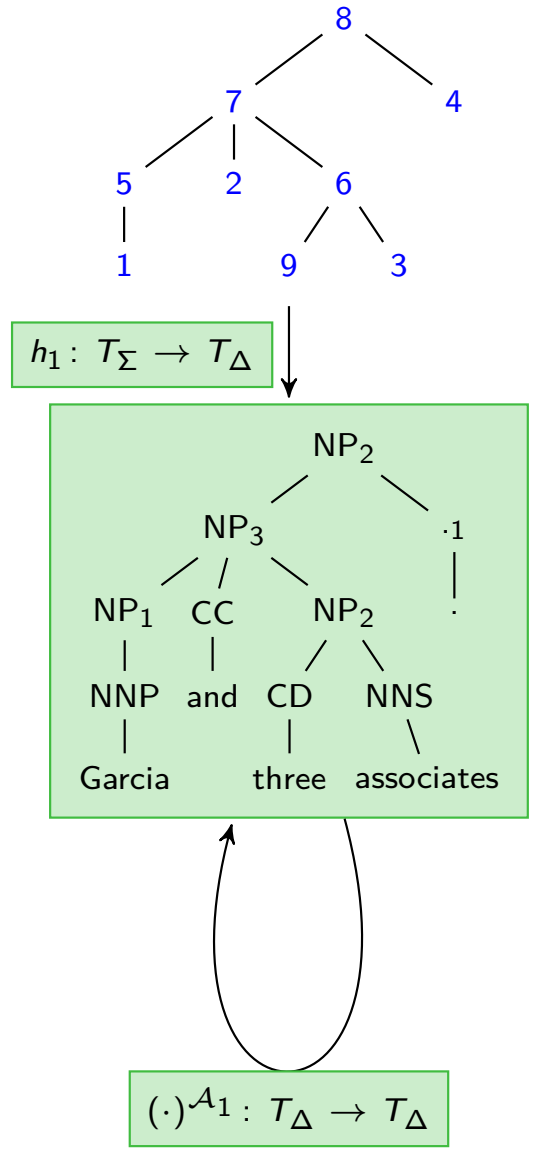




IRTG

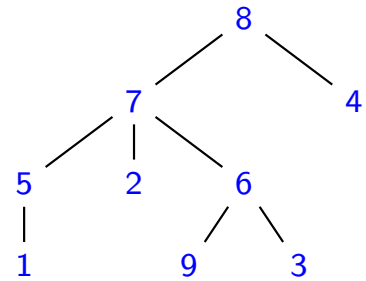
- wRTG (shown as hypergraph)
- tree homomorphism h_1
- algebra $\mathcal{A}_1 = \langle T_\Delta, (\cdot)^{\mathcal{A}_1} \rangle$

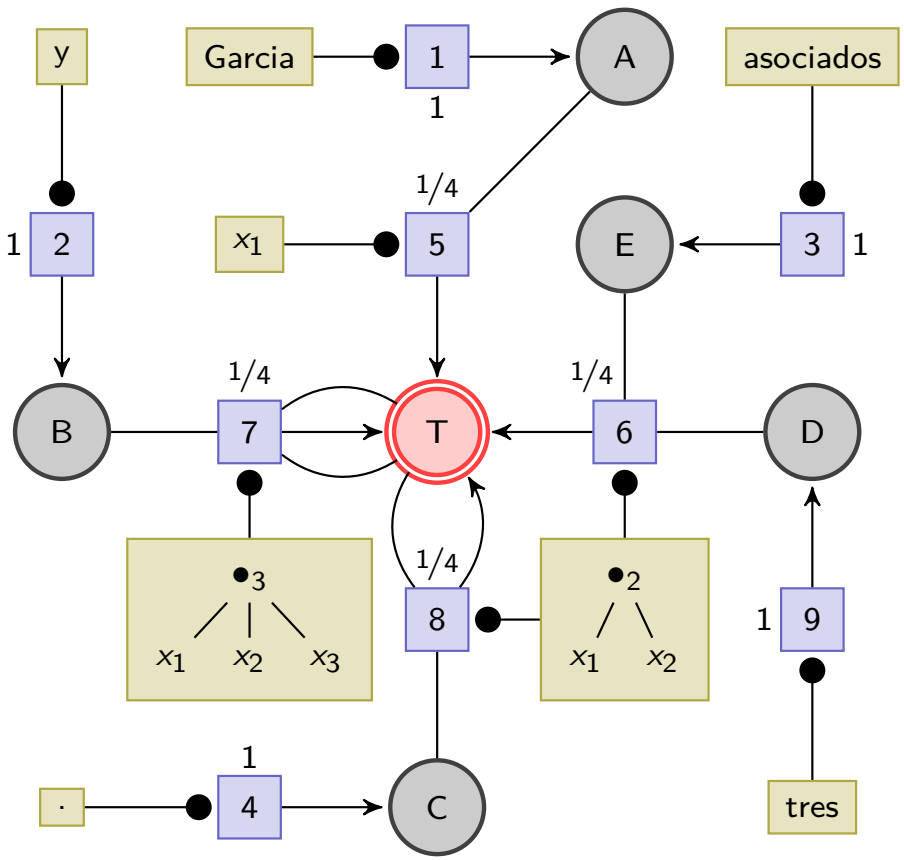




IRTG

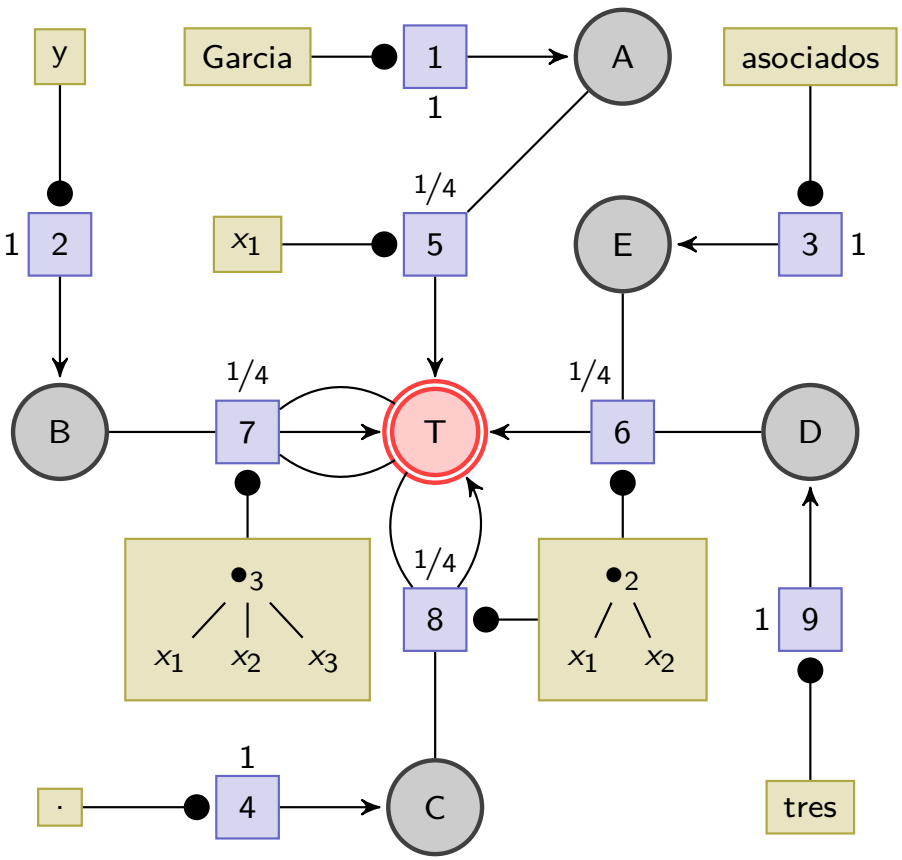
- wRTG (shown as hypergraph)
- tree homomorphisms h_1, h_2
- algebra $\mathcal{A}_1 = \langle T_\Delta, (\cdot)^{\mathcal{A}_1} \rangle$

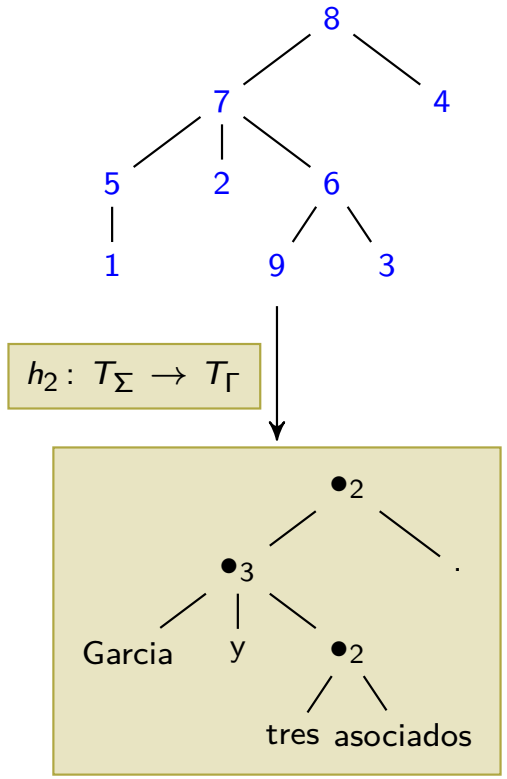




IRTG

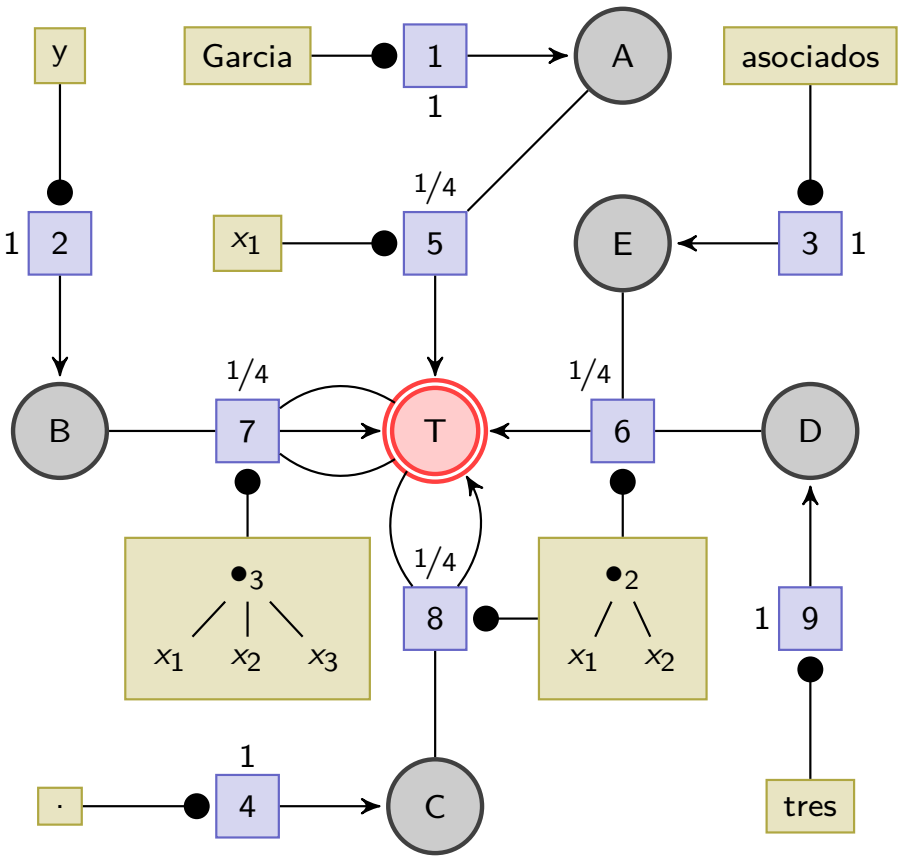
- wRTG (shown as hypergraph)
- tree homomorphisms h_1, h_2
- algebra $\mathcal{A}_1 = \langle T_\Delta, (\cdot)^{\mathcal{A}_1} \rangle$

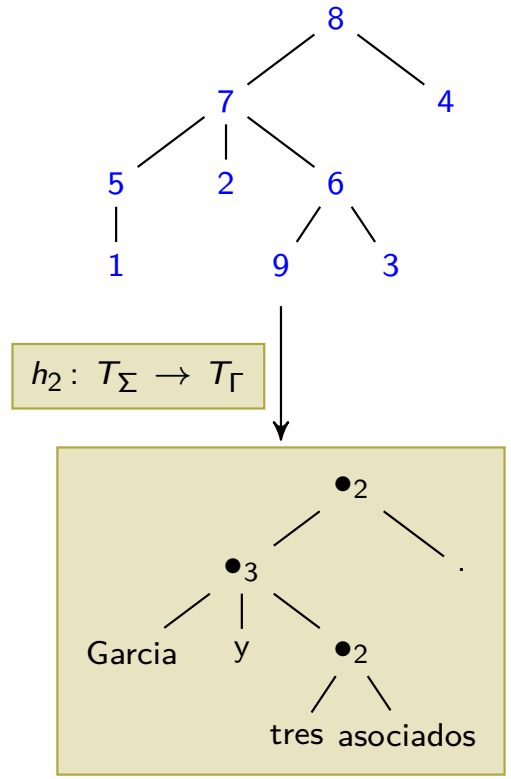




IRTG

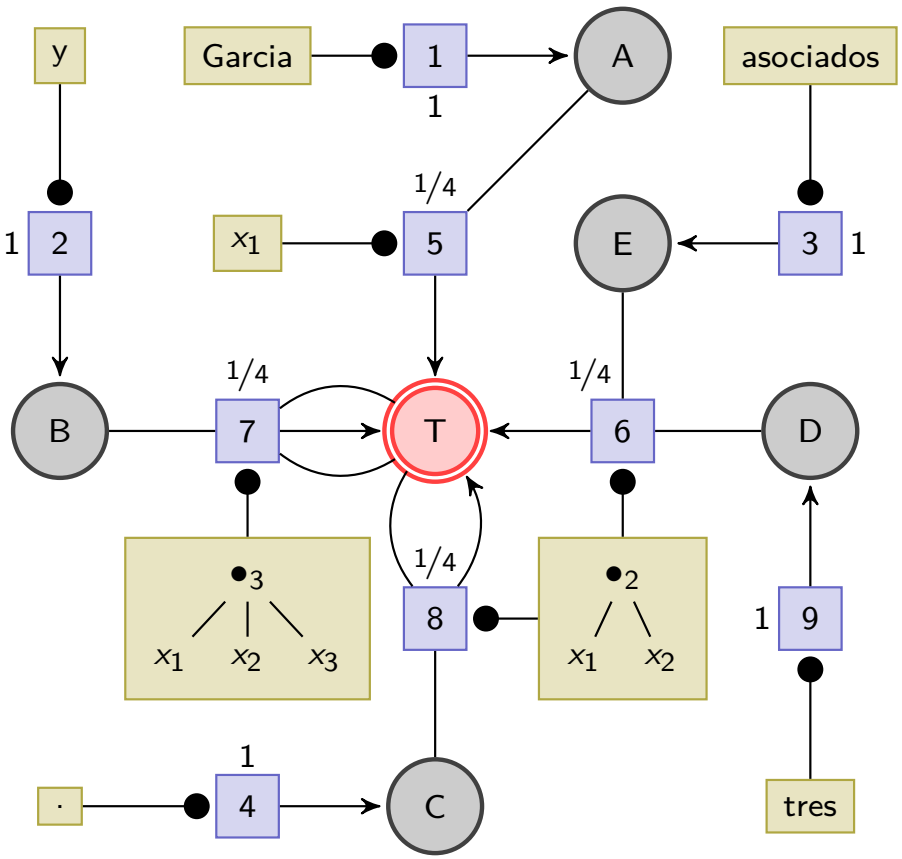
- wRTG (shown as hypergraph)
- tree homomorphisms h_1, h_2
- algebras $\mathcal{A}_1 = \langle T_\Delta, (\cdot)^{\mathcal{A}_1} \rangle$, $\mathcal{A}_2 = \langle \Gamma^*, (\cdot)^{\mathcal{A}_2} \rangle$

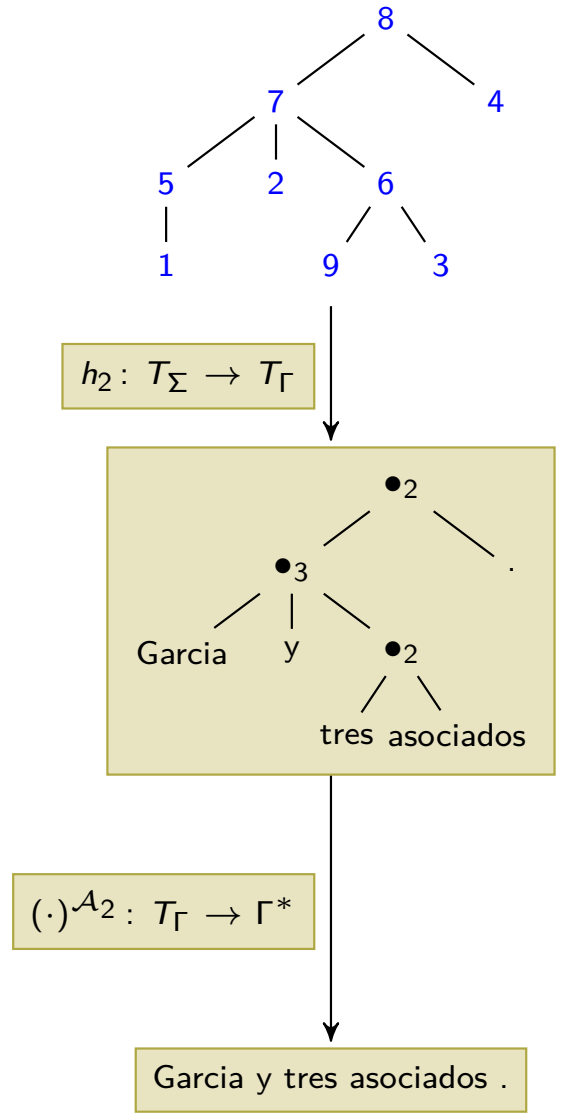




IRTG

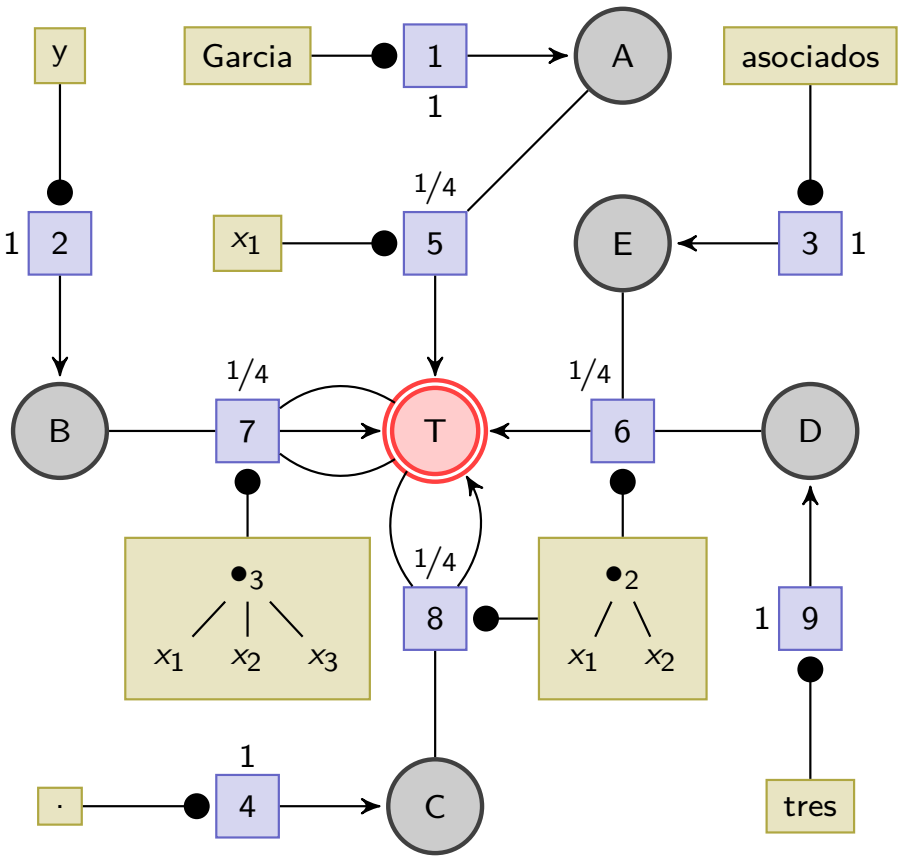
- wRTG (shown as hypergraph)
- tree homomorphisms h_1, h_2
- algebras $\mathcal{A}_1 = \langle T_\Delta, (\cdot)^{\mathcal{A}_1} \rangle$, $\mathcal{A}_2 = \langle \Gamma^*, (\cdot)^{\mathcal{A}_2} \rangle$

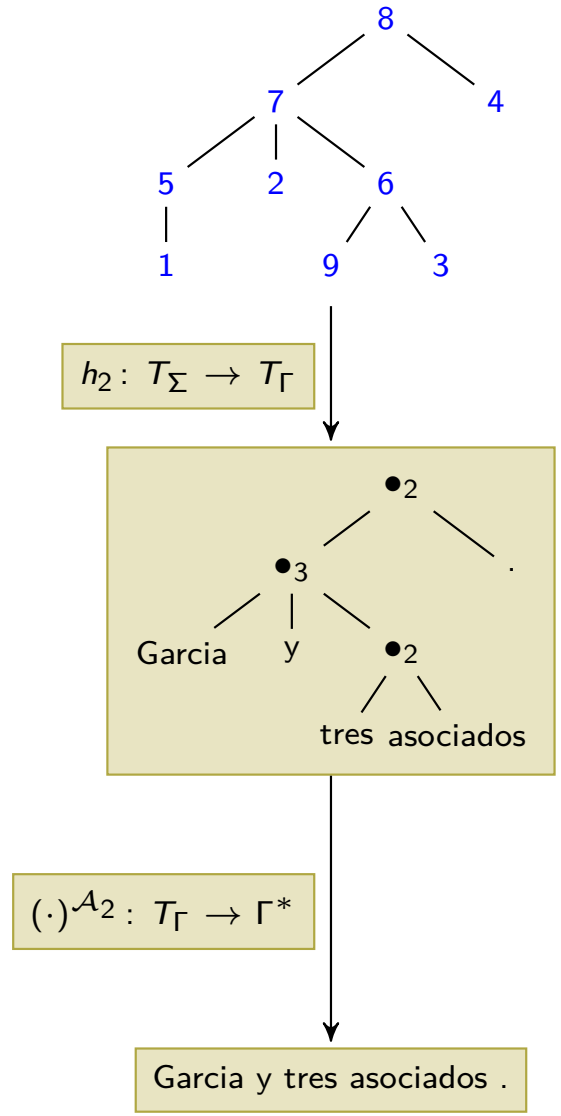




IRTG

- wRTG (shown as hypergraph)
- tree homomorphisms h_1, h_2
- algebras $\mathcal{A}_1 = \langle T_\Delta, (\cdot)^{\mathcal{A}_1} \rangle$, $\mathcal{A}_2 = \langle \Gamma^*, (\cdot)^{\mathcal{A}_2} \rangle$
- weighted language over $T_\Delta \times \Gamma^*$





Outline

n-Gram Model

example: 3-gram model ($n = 3$)

n-Gram Model

example: 3-gram model ($n = 3$)

- scores a sequence of symbols

$w =$ Garcia | y | tres | asociados | .

n-Gram Model

example: 3-gram model ($n = 3$)

- scores a sequence of symbols

$w =$ Garcia | y | tres | asociados | .

$N(w) =$

n-Gram Model

example: 3-gram model ($n = 3$)

- scores a sequence of symbols
- assigns weight to every $\underbrace{\text{sequence of 3 symbols}}_{= \text{3-gram}}$

$w =$ Garcia | y | tres | asociados | .

$N(w) =$ $2/5$

n-Gram Model

example: 3-gram model ($n = 3$)

- scores a sequence of symbols
- assigns weight to every $\underbrace{\text{sequence of 3 symbols}}_{= \text{3-gram}}$

$w =$ Garcia | y | tres | asociados | .

$N(w) =$ $2/5$ $1/3$

n-Gram Model

example: 3-gram model ($n = 3$)

- scores a sequence of symbols
- assigns weight to every $\underbrace{\text{sequence of 3 symbols}}_{= \text{3-gram}}$

$w =$ Garcia | y | tres | asociados | .

$N(w) =$ $2/5$ $1/3$ $1/4$

n-Gram Model

example: 3-gram model ($n = 3$)

- scores a sequence of symbols
- assigns weight to every $\underbrace{\text{sequence of 3 symbols}}_{= \text{3-gram}}$
- weights are multiplied

$$w = \boxed{\text{Garcia}}\ \boxed{\text{y}}\ \boxed{\text{tres}}\ \boxed{\text{asociados}}\ \boxed{\text{.}}$$

$$N(w) = {}^{2}/_{5} \cdot {}^{1}/_{3} \cdot {}^{1}/_{4} = {}^{1}/_{30}$$

Outline

Component Product

$$h_{p_1,p_2}(f) = \text{argmax}_{e \in E} \underbrace{p_1}_{\text{by IRTG}}(f \mid e) \cdot \underbrace{p_2}_{\text{by } n\text{-gram model}}(e)$$

Component Product

$$h_{p_1,p_2}(f) = \text{argmax}_{e \in E} \underbrace{p_1}_{\text{by IRTG}}(f \mid e) \cdot \underbrace{p_2}_{\text{by } n\text{-gram model}}(e)$$

$$h_{\mathcal{L},N}(f) = \text{argmax}_{e \in E}\, \mathcal{L}(\langle e, f \rangle) \cdot N(e)$$

Component Product

$$h_{p_1,p_2}(f) = \text{argmax}_{e \in E} \underbrace{p_1}_{\text{by IRTG}}(f \mid e) \cdot \underbrace{p_2}_{\text{by } n\text{-gram model}}(e)$$

$$\begin{aligned} h_{\mathcal{L},N}(f) &= \text{argmax}_{e \in E}\, \mathcal{L}(\langle e, f \rangle) \cdot N(e) \\ &= \text{argmax}_{e \in E} \underbrace{(\mathcal{L} \odot_1 N)}_{\text{component product}}(\langle e, f \rangle) \end{aligned}$$

Component Product

$$h_{p_1,p_2}(f) = \text{argmax}_{e \in E} \underbrace{p_1}_{\text{by IRTG}}(f \mid e) \cdot \underbrace{p_2}_{\text{by } n\text{-gram model}}(e)$$

$$\begin{aligned} h_{\mathcal{L},N}(f) &= \text{argmax}_{e \in E}\, \mathcal{L}(\langle e, f \rangle) \cdot N(e) \\ &= \text{argmax}_{e \in E} \underbrace{(\mathcal{L} \odot_1 N)}_{\text{component product}}(\langle e, f \rangle) \end{aligned}$$

$$(\mathcal{L} \odot_i N)(\langle w_1, \ldots, w_k \rangle) = \mathcal{L}(\langle w_1, \ldots, w_k \rangle) \cdot N(w_i)$$

Results

Theorem (main result)

$$\mathit{IRTG} \odot_i \mathit{n\text{-}gram\ model} \subseteq \mathit{IRTG}$$

Results

Theorem (main result)

$$\mathit{IRTG} \odot_i \mathit{n\text{-}gram\ model} \subseteq \mathit{IRTG}$$

Lemma (*n*-gram wRTG)

$$(\cdot)^{\mathcal{A}}; \mathit{n\text{-}gram\ model} \subseteq \mathit{wRTG}$$

Results

Theorem (main result)

$$\mathit{IRTG} \odot_i \mathit{n\text{-}gram\ model} \subseteq \mathit{IRTG}$$

Lemma (*n*-gram wRTG)

$$(\cdot)^{\mathcal{A}}; \mathit{n\text{-}gram\ model} \subseteq \mathit{wRTG}$$

Lemma (inverse homomorphism wRTG)

$$\mathit{ln\text{-}Hom}^{-1}(\mathit{wRTG}) \subseteq \mathit{wRTG}$$

Results

Theorem (main result)

$$\mathit{IRTG} \odot_i \textit{n-gram model} \subseteq \mathit{IRTG}$$

Lemma (*n*-gram wRTG)

$$(\cdot)^{\mathcal{A}}; \textit{n-gram model} \subseteq \mathit{wRTG}$$

Lemma (inverse homomorphism wRTG)

$$\mathit{ln\text{-}Hom}^{-1}(\mathit{wRTG}) \subseteq \mathit{wRTG}$$

Lemma (Hadamard product)

$$\mathit{wRTG} \odot \mathit{wRTG} \subseteq \mathit{wRTG}$$

Results

Theorem (main result)

$$IRTG \odot_i \textit{n-gram model} \subseteq IRTG$$

Lemma (*n*-gram wRTG)

$$(\cdot)^{\mathcal{A}}; \textit{n-gram model} \subseteq \textit{wRTG}$$

Lemma (inverse homomorphism wRTG)

$$\textit{In-Hom}^{-1}(\textit{wRTG}) \subseteq \textit{wRTG}$$

Lemma (Hadamard product)

$$\textit{wRTG} \odot \textit{wRTG} \subseteq \textit{wRTG}$$

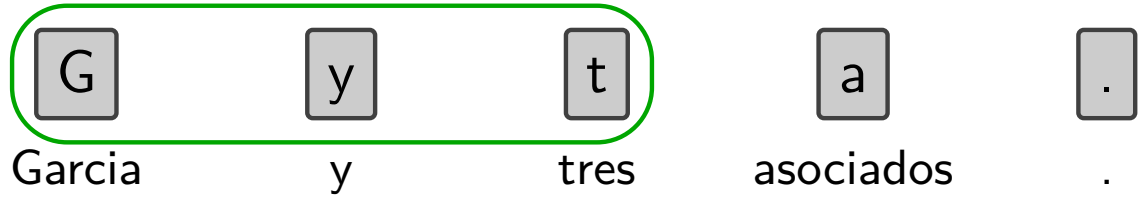
Outline

n-Gram wRTG

$w =$ 

$N(w) =$

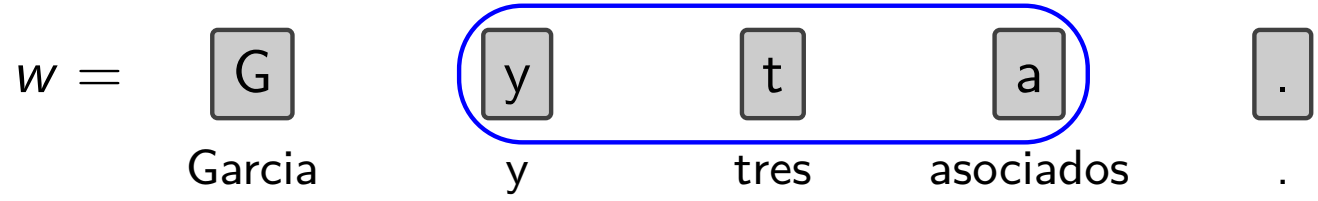
n-Gram wRTG

$w =$ 

$N(w) =$

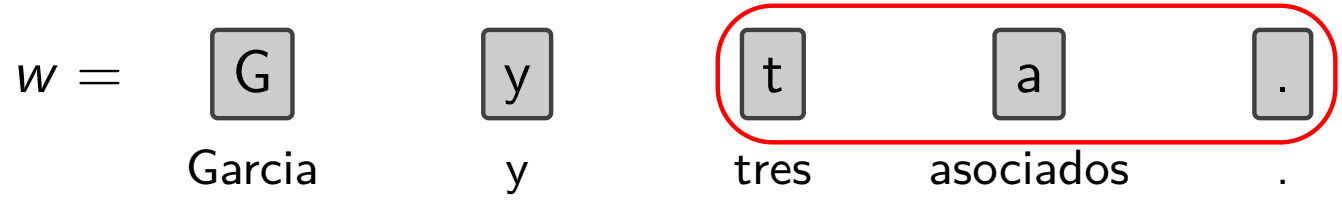
$N(\text{Gyt})$
$2/5$

n-Gram wRTG



$$N(w) = \underset{2/5}{N(\text{Gyt})} \cdot \underset{1/3}{N(\text{yta})}$$

n-Gram wRTG



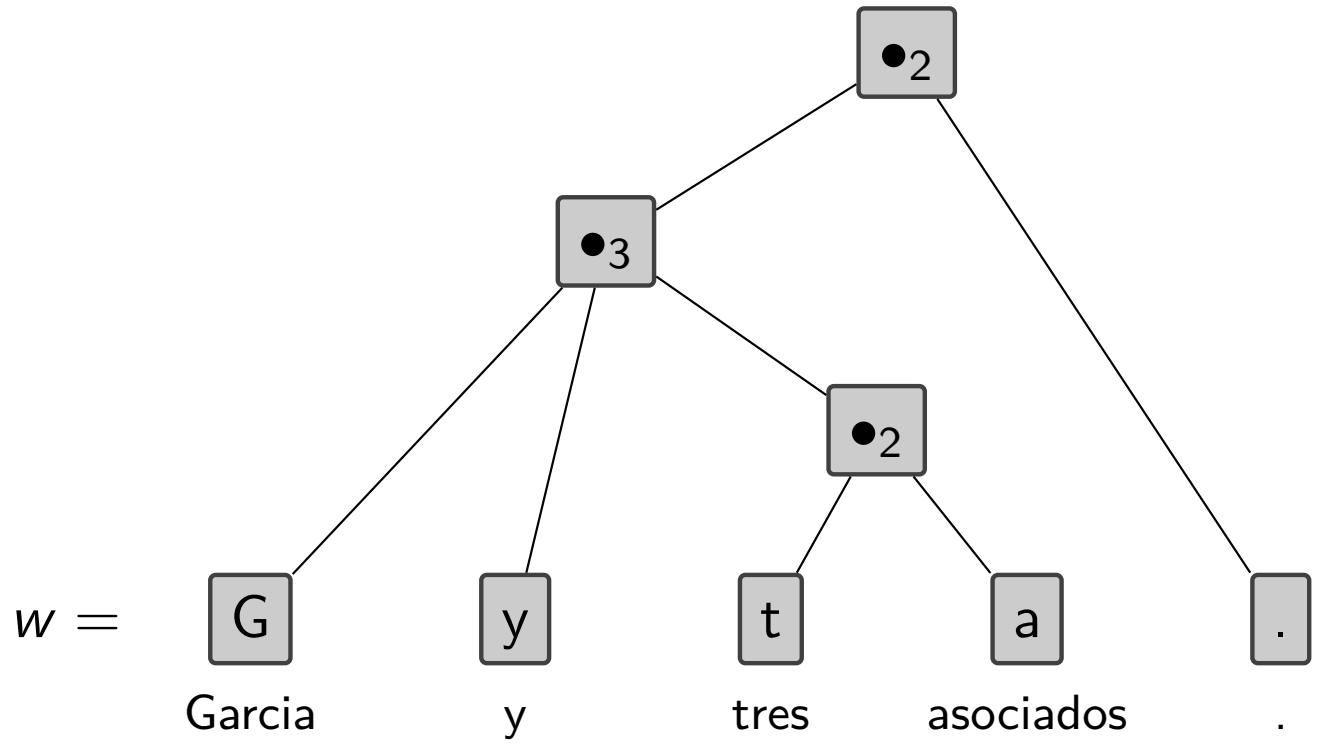
$$N(w) = \underset{N(\text{Gyt})}{2/5} \cdot \underset{N(\text{yta})}{1/3} \cdot \underset{N(\text{ta.})}{1/4}$$

n-Gram wRTG

$$w = \underset{\text{Garcia}}{\boxed{\text{G}}} \quad \underset{\text{y}}{\boxed{\text{y}}} \quad \underset{\text{tres}}{\boxed{\text{t}}} \quad \underset{\text{asociados}}{\boxed{\text{a}}} \quad \underset{\text{.}}{\boxed{\text{.}}}$$

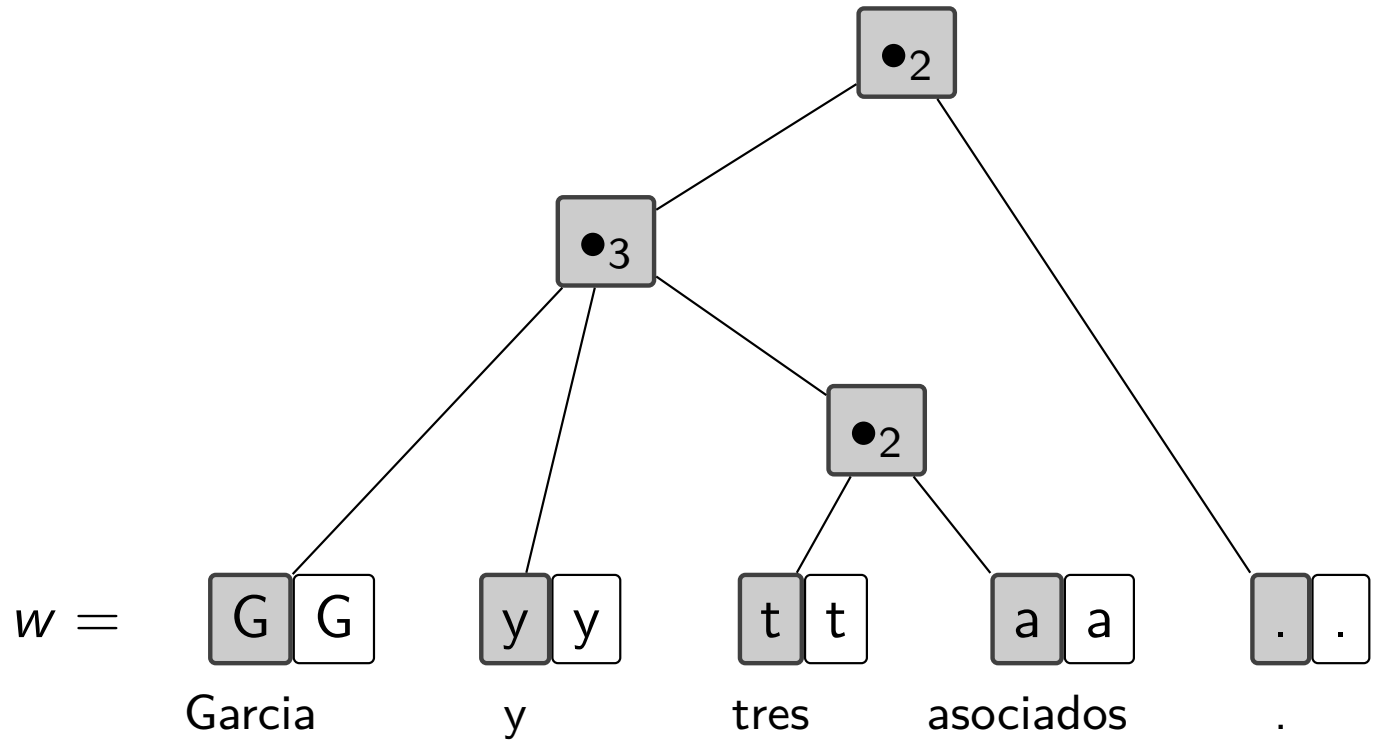
$$N(w) = \underset{N(\text{Gyt})}{2/5} \cdot \underset{N(\text{yta})}{1/3} \cdot \underset{N(\text{ta.})}{1/4} = 1/30$$

n-Gram wRTG



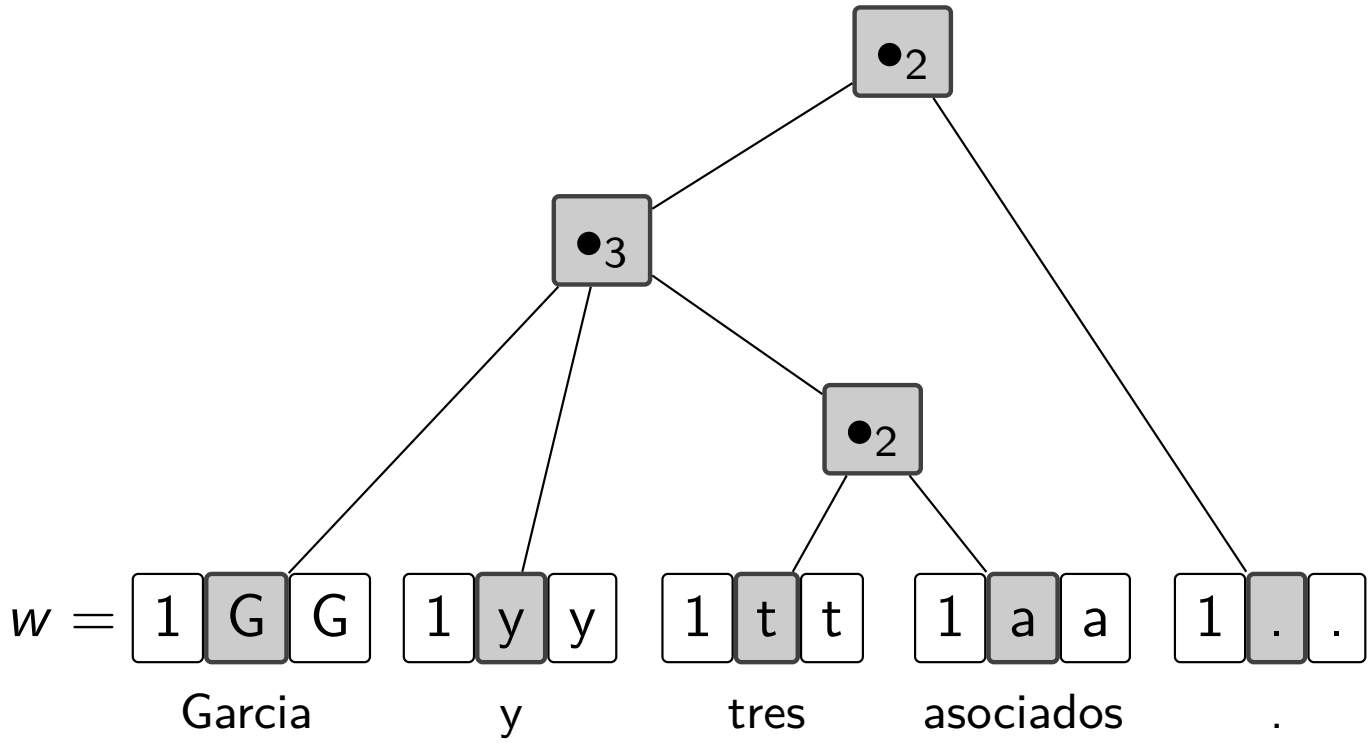
$$N(w) = \underbrace{2/5}_{N(\text{Gyt})} \cdot \underbrace{1/3}_{N(\text{yta})} \cdot \underbrace{1/4}_{N(\text{ta.})} = 1/30$$

n-Gram wRTG



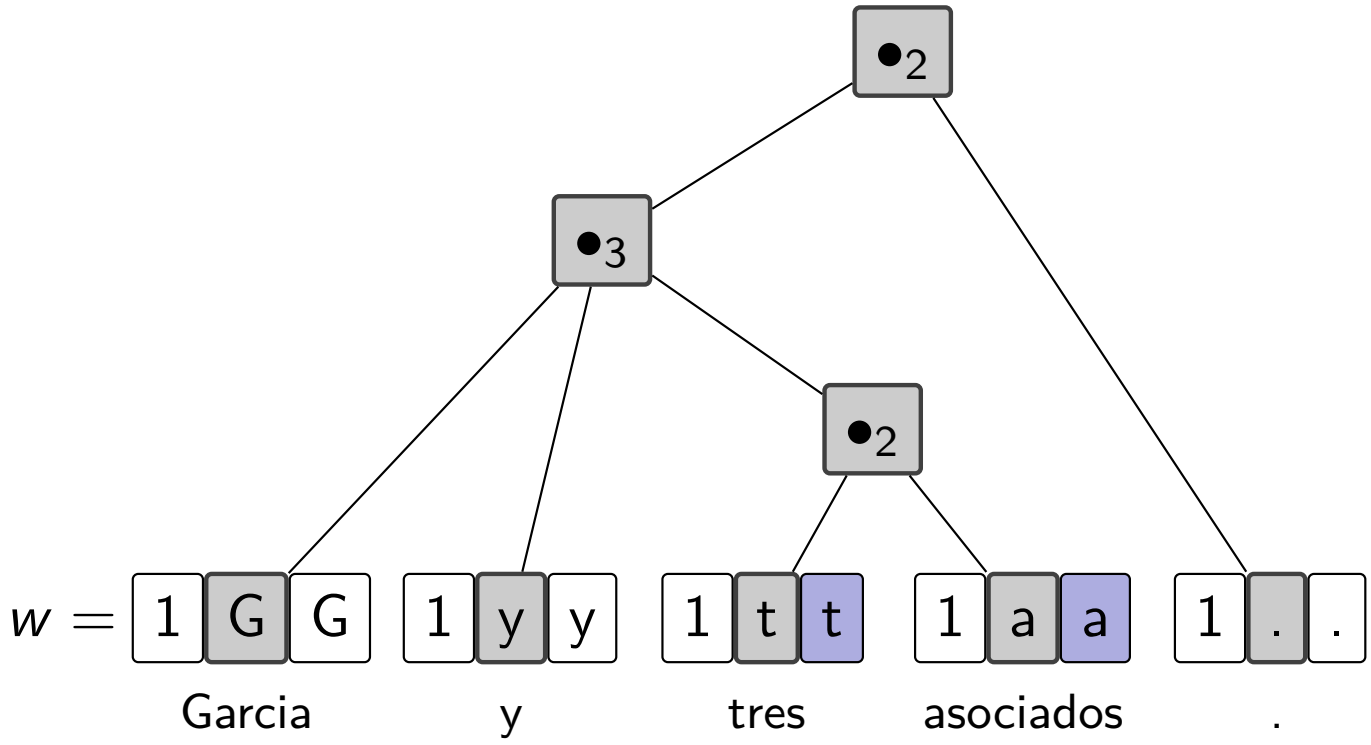
$$N(w) = \underbrace{2/5}_{N(\text{Gyt})} \cdot \underbrace{1/3}_{N(\text{yta})} \cdot \underbrace{1/4}_{N(\text{ta.})} = 1/30$$

n-Gram wRTG



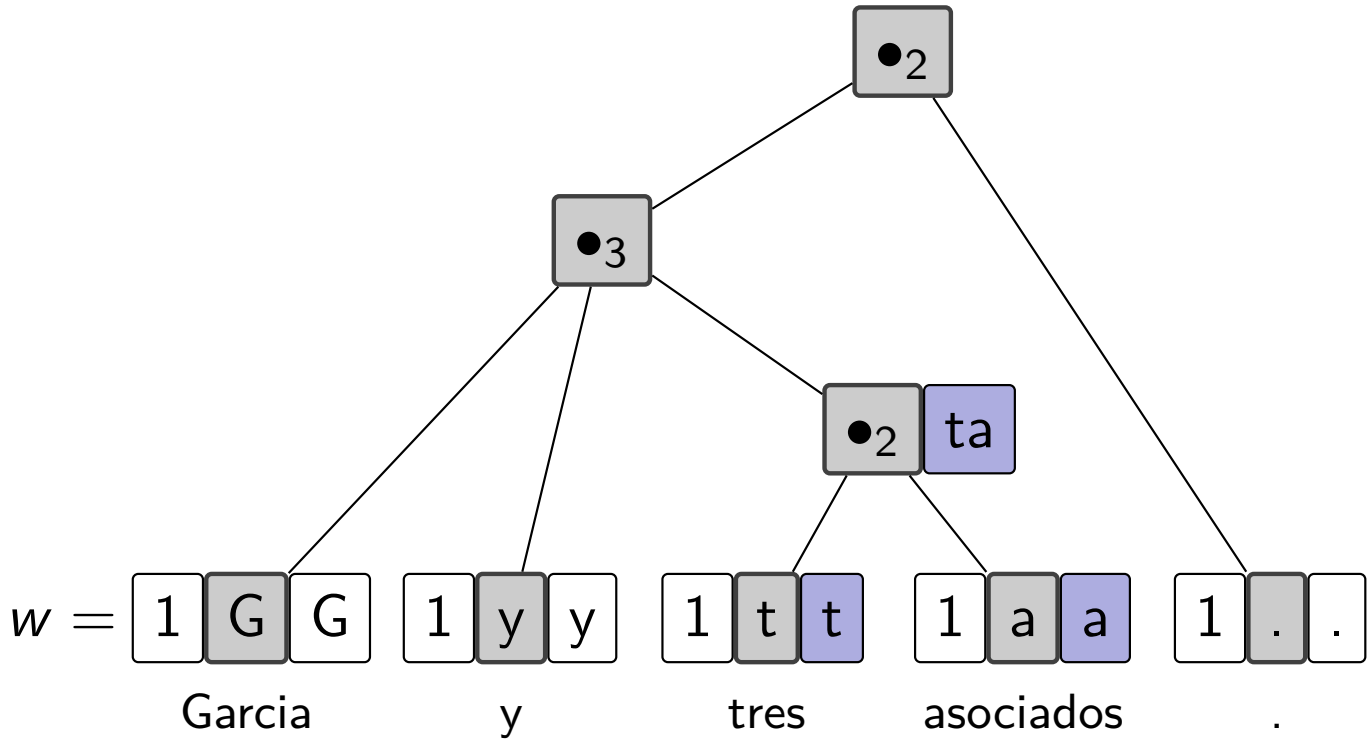
$$N(w) = \underbrace{2/5}_{N(\text{Gyt})} \cdot \underbrace{1/3}_{N(\text{yta})} \cdot \underbrace{1/4}_{N(\text{ta.})} = 1/30$$

n-Gram wRTG



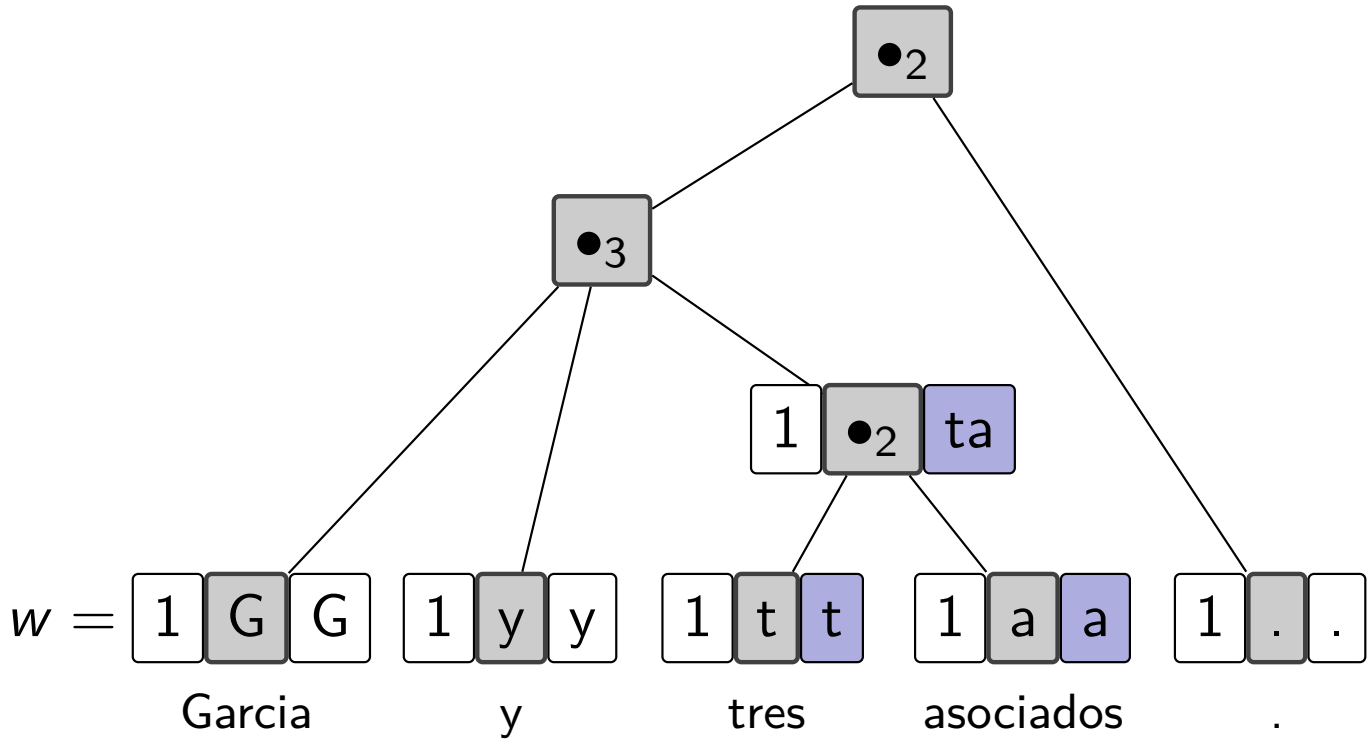
$$N(w) = \underbrace{2/5}_{N(\text{Gyt})} \cdot \underbrace{1/3}_{N(\text{yta})} \cdot \underbrace{1/4}_{N(\text{ta.})} = 1/30$$

n-Gram wRTG



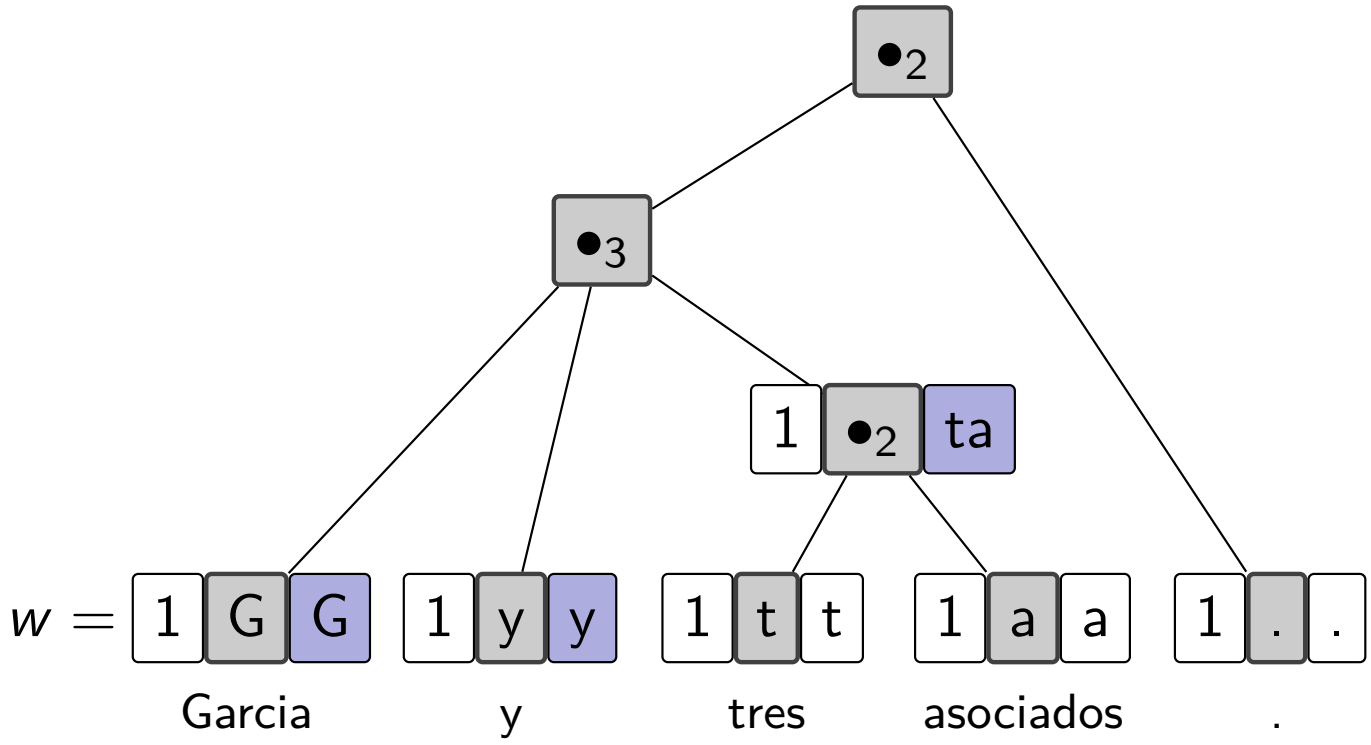
$$N(w) = \underbrace{2/5}_{N(\text{Gyt})} \cdot \underbrace{1/3}_{N(\text{yta})} \cdot \underbrace{1/4}_{N(\text{ta.})} = 1/30$$

n-Gram wRTG



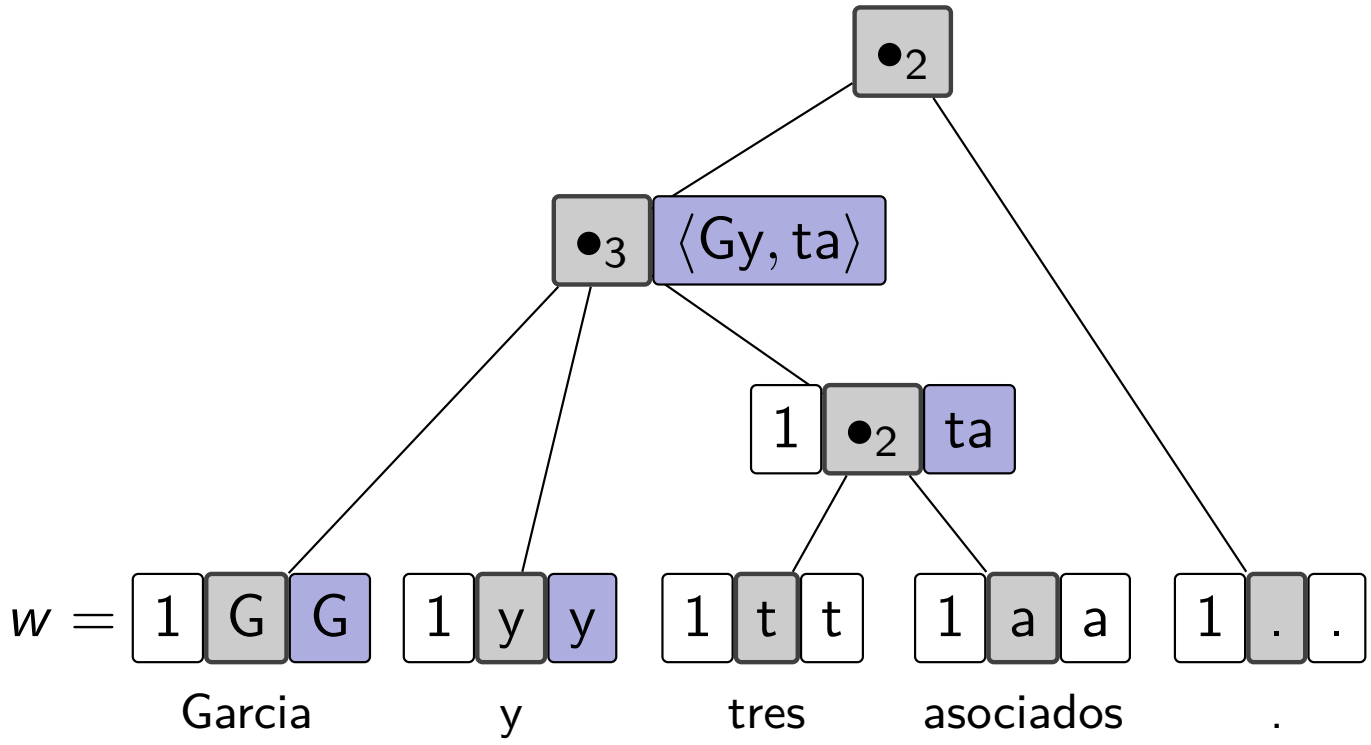
$$N(w) = \underbrace{2/5}_{N(\text{Gyt})} \cdot \underbrace{1/3}_{N(\text{yta})} \cdot \underbrace{1/4}_{N(\text{ta.})} = 1/30$$

n-Gram wRTG



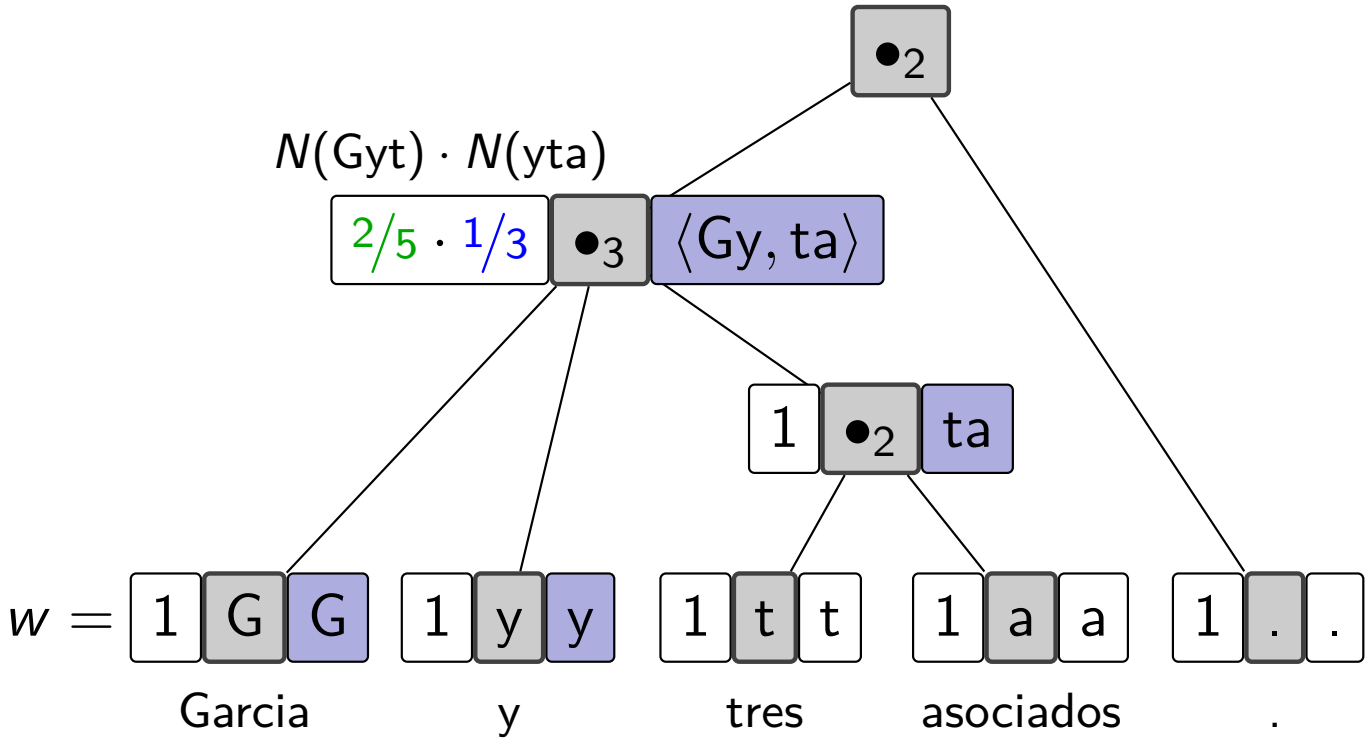
$$N(w) = \underbrace{2/5}_{N(\text{Gyt})} \cdot \underbrace{1/3}_{N(\text{yta})} \cdot \underbrace{1/4}_{N(\text{ta.})} = 1/30$$

n-Gram wRTG



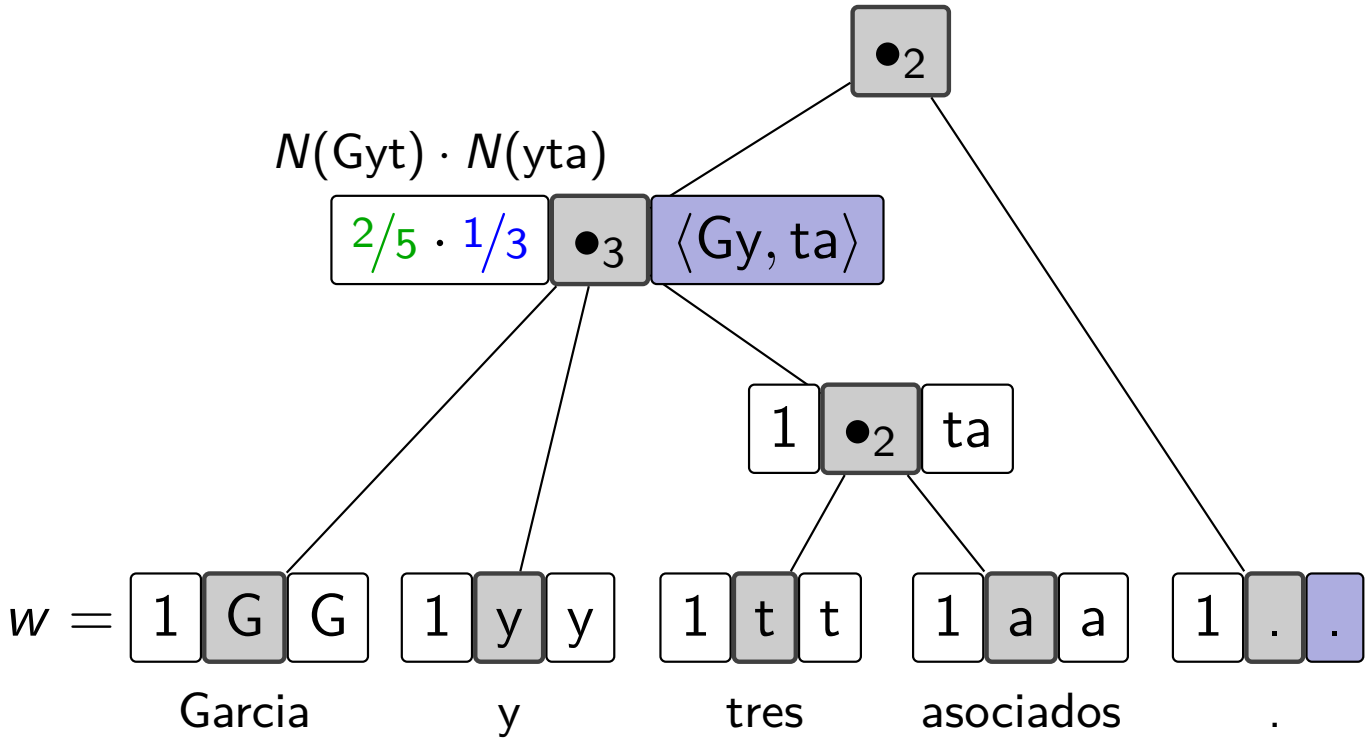
$$N(w) = \underbrace{2/5}_{N(\text{Gyt})} \cdot \underbrace{1/3}_{N(\text{yta})} \cdot \underbrace{1/4}_{N(\text{ta.})} = 1/30$$

n-Gram wRTG



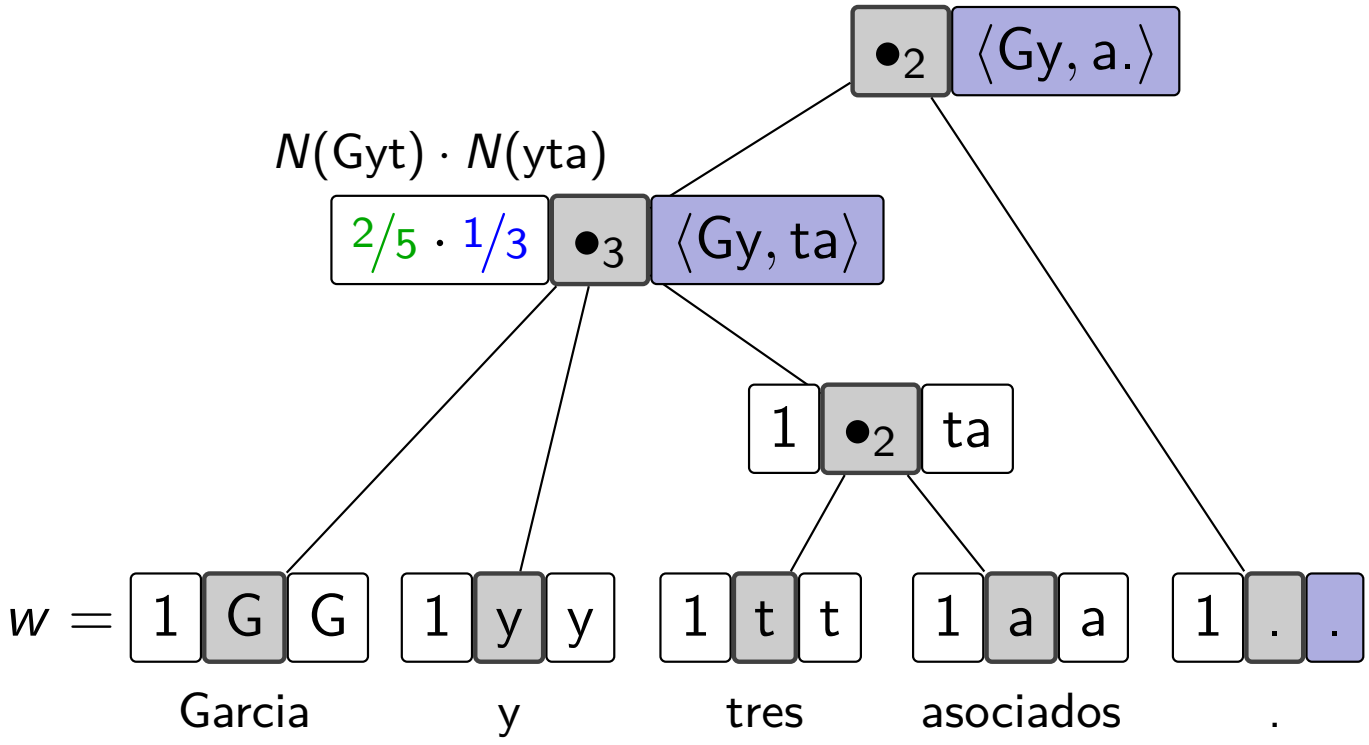
$$N(w) = \overbrace{2/5}^{N(\text{Gyt})} \cdot \overbrace{1/3}^{N(\text{yta})} \cdot \overbrace{1/4}^{N(\text{ta.})} = 1/30$$

n-Gram wRTG



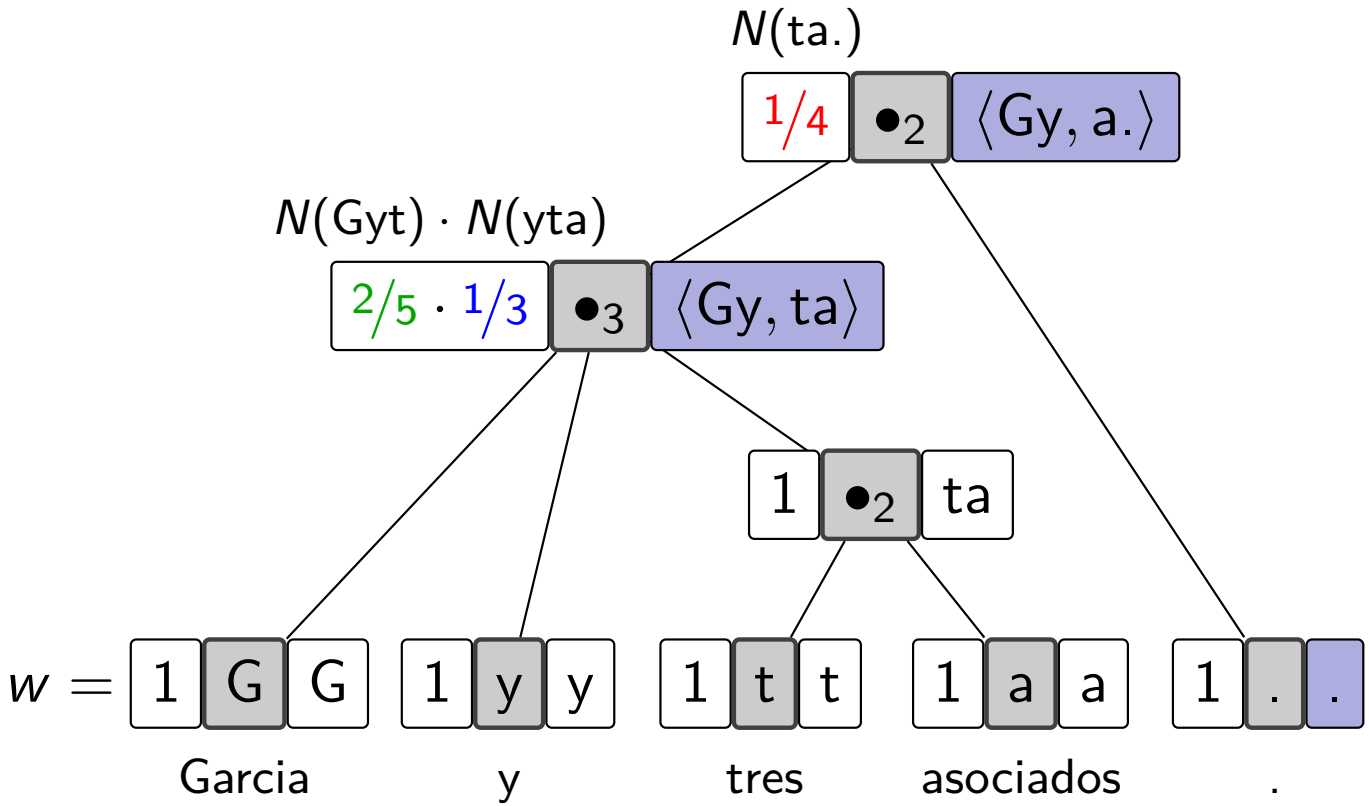
$$N(w) = \underbrace{2/5}_{N(\text{Gyt})} \cdot \underbrace{1/3}_{N(\text{yta})} \cdot \underbrace{1/4}_{N(\text{ta.})} = 1/30$$

n-Gram wRTG



$$N(w) = \underbrace{2/5}_{N(\text{Gyt})} \cdot \underbrace{1/3}_{N(\text{yta})} \cdot \underbrace{1/4}_{N(\text{ta.})} = 1/30$$

n-Gram wRTG



$$N(w) = \underbrace{\boxed{2/5}}_{N(\text{Gyt})} \cdot \underbrace{\boxed{1/3}}_{N(\text{yta})} \cdot \underbrace{\boxed{1/4}}_{N(\text{ta.})} = 1/30$$

Outline

Summary

$$h_{p_1,p_2}(f) = \text{argmax}_{e \in E} \underbrace{p_1}_{\text{by IRTG}}(f \mid e) \cdot \underbrace{p_2}_{\text{by } n\text{-gram model}}(e)$$

Summary

$$h_{p_1,p_2}(f) = \text{argmax}_{e \in E} \underbrace{p_1}_{\text{by IRTG}}(f \mid e) \cdot \underbrace{p_2}_{\text{by } n\text{-gram model}}(e)$$

$$\text{IRTG} \odot_i n\text{-gram model} \subseteq \text{IRTG}$$

Summary

$$h_{p_1,p_2}(f) = \text{argmax}_{e \in E} \underbrace{p_1}_{\text{by IRTG}}(f \mid e) \cdot \underbrace{p_2}_{\text{by } n\text{-gram model}}(e)$$

$$\text{IRTG} \odot_i n\text{-gram model} \subseteq \text{IRTG}$$

decompose result:

Summary

$$h_{p_1,p_2}(f) = \text{argmax}_{e \in E} \underbrace{p_1}_{\text{by IRTG}}(f \mid e) \cdot \underbrace{p_2}_{\text{by } n\text{-gram model}}(e)$$

IRTG $\odot_i$ n-gram model $\subseteq$ IRTG

decompose result:

1. $(\cdot)^{\mathcal{A}}$; n-gram model $\subseteq$ wRTG
   - calculate N when possible
   - store unprocessed strings and margins in states

Summary

$$h_{p_1,p_2}(f) = \text{argmax}_{e \in E} \underbrace{p_1}_{\text{by IRTG}}(f \mid e) \cdot \underbrace{p_2}_{\text{by } n\text{-gram model}}(e)$$

IRTG $\odot_i$ n-gram model $\subseteq$ IRTG

decompose result:

1. $(\cdot)^{\mathcal{A}}$; n-gram model $\subseteq$ wRTG
   - calculate N when possible
   - store unprocessed strings and margins in states
2. In-Hom^{-1}(wRTG) $\subseteq$ wRTG
   - usual construction, not covered here

Summary

$$h_{p_1,p_2}(f) = \text{argmax}_{e \in E} \underbrace{p_1}_{\text{by IRTG}}(f \mid e) \cdot \underbrace{p_2}_{\text{by } n\text{-gram model}}(e)$$

IRTG $\odot_i$ n-gram model $\subseteq$ IRTG

decompose result:

1. $(\cdot)^{\mathcal{A}}$; n-gram model $\subseteq$ wRTG
   - calculate N when possible
   - store unprocessed strings and margins in states
2. ln-Hom^{-1}(wRTG) $\subseteq$ wRTG
   - usual construction, not covered here
3. wRTG $\odot$ wRTG $\subseteq$ wRTG
   - usual construction, not covered here

References

J. Berstel and C. Reutenauer. "Recognizable formal power series on trees". In: *Theoretical Computer Science* 18.2 (1982), pp. 115–148. ISSN: 0304-3975.

David Chiang. "Hierarchical Phrase-Based Translation". In: *Comput. Linguist.* 33.2 (June 2007), pp. 201–228. ISSN: 0891-2017.

Zoltán Fülöp, Andreas Maletti, and Heiko Vogler. *Weighted Extended Tree Transducers*. 2010.