Decoding for SMT

Wilker Aziz

Universiteit van Amsterdam
w.aziz@uva.nl

April 19, 2016

Content

Model of translational equivalences

Describes the process of generating translations of a given input

- constrains and characterises
  the set of possible translation derivations

Model of translational equivalences

Describes the process of generating translations of a given input

- constrains and characterises
  the set of possible translation derivations

Phrase-based MT

we observe an input, segment it into phrases, permute the phrases into target language word-order, and finally, translate segments independently

Model of translational equivalences

Describes the process of generating translations of a given input

- constrains and characterises
  the set of possible translation derivations

Phrase-based MT

we observe an input, segment it into phrases, permute the phrases into target language word-order, and finally, translate segments independently

Hierarchical MT

we parse the input with a CFG, then translate (using synchronous rules) each and every edge independently

CFGs and FSAs

Compactly represent the set of translations

- keep the representation cost a tractable (polynomial) function of the input length

Phrase-based MT $O(n^2 2^d)$

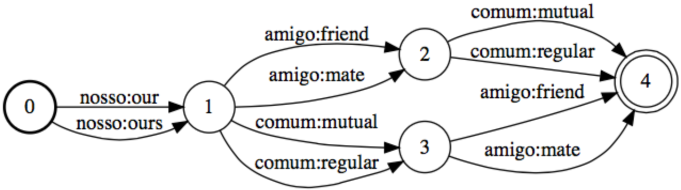
Hierarchical MT $O(n^3)$

Independence assumptions

Translation rules (flat or CFG) are applied independently

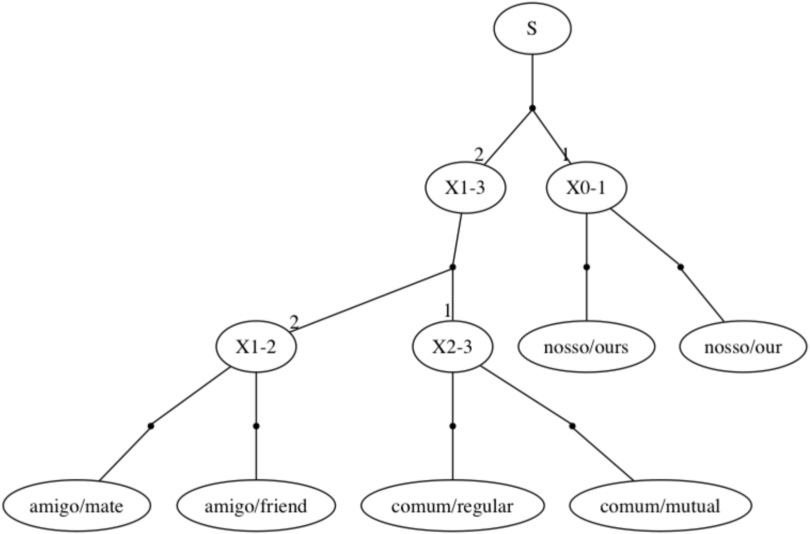
Independence assumptions

Translation rules (flat or CFG) are applied independently



Independence assumptions

Translation rules (flat or CFG) are applied independently



Directed B-hypergraphs

A hypergraph $\langle V, E \rangle$ consists of

- a set of nodes V
- a set of edges E
- an edge e has
  - a head node $\mathrm{head}(e) \in V$
  - a tail $\mathrm{tail}(e) \in V^*$ (sequence of nodes)

Directed B-hypergraphs

A hypergraph $\langle V, E \rangle$ consists of

- a set of nodes V
- a set of edges E
- an edge e has
  - a head node $\mathrm{head}(e) \in V$
  - a tail $\mathrm{tail}(e) \in V^*$ (sequence of nodes)

CFGs

- nonterminal $\rightarrow$ node
- terminal $\rightarrow$ terminal node
- rule $\rightarrow$ edge
- LHS $\rightarrow$ head
- RHS $\rightarrow$ tail

Directed B-hypergraphs

A hypergraph $\langle V, E \rangle$ consists of

- a set of nodes V
- a set of edges E
- an edge e has
  - a head node $\mathrm{head}(e) \in V$
  - a tail $\mathrm{tail}(e) \in V^*$ (sequence of nodes)

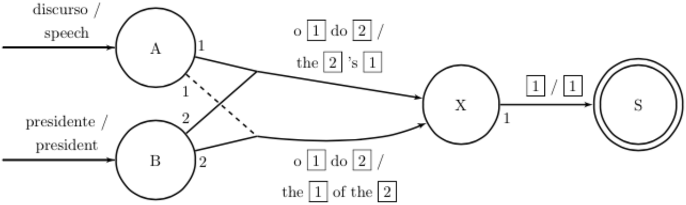
CFGs

- nonterminal $\rightarrow$ node
- terminal $\rightarrow$ terminal node
- rule $\rightarrow$ edge
- LHS $\rightarrow$ head
- RHS $\rightarrow$ tail

FSAs

- state $\rightarrow$ node
- symbol $\rightarrow$ terminal node
- transition $\rightarrow$ edge
- origin $\rightarrow$ tail node
- destination $\rightarrow$ head

A forest as a hypergraph



(a) Hypergraph

LHS	RHS_i	RHS_o
S →	X	$\boxed{1}$
X →	o A do B	the $\boxed{2}$'s $\boxed{1}$
X →	o A do B	the $\boxed{1}$ of the $\boxed{2}$
A →	discurso	speech
B →	presidente	president

(b) Synchronous rules

Weighted sets

A weighted set $\langle \mathcal{D}, \omega \rangle$ consists of

- a set of structures (e.g. hyperpaths/derivations)
- a function $w : \mathcal{D} \to \mathcal{K}$

Let us focus on weighted sets whose weight functions factorise

$$w(\mathbf{d}) = \bigotimes_{e \in \mathbf{d}} w(e)$$

Often the structure is just a means to an end (the yield)

$$w(\mathbf{y}) = \bigoplus_{\mathbf{d} \in \mathcal{D}_{\mathbf{y}}} w(\mathbf{d})$$

Semirings

An algebraic structure $\mathcal{K} = \langle \mathbb{K}, \oplus, \otimes, \bar{0}, \bar{1} \rangle$

Semirings

An algebraic structure $\mathcal{K} = \langle \mathbb{K}, \oplus, \otimes, \bar{0}, \bar{1} \rangle$

- $\mathbb{K}$ is a set (e.g. $\mathbb{N}, \mathbb{R}, \{0, 1\}$)

Semirings

An algebraic structure $\mathcal{K} = \langle \mathbb{K}, \oplus, \otimes, \bar{0}, \bar{1} \rangle$

- $\mathbb{K}$ is a set (e.g. $\mathbb{N}, \mathbb{R}, \{0, 1\}$)
- $\oplus$ and $\otimes$ are binary operators

Semirings

An algebraic structure $\mathcal{K} = \langle \mathbb{K}, \oplus, \otimes, \bar{0}, \bar{1} \rangle$

- $\mathbb{K}$ is a set (e.g. $\mathbb{N}, \mathbb{R}, \{0, 1\}$)
- $\oplus$ and $\otimes$ are binary operators
- $\oplus$ is commutative and has identity $\bar{0}$
  $a \oplus b = b \oplus a$ and $\bar{0} \oplus a = a \oplus \bar{0} = a$

Semirings

An algebraic structure $\mathcal{K} = \langle \mathbb{K}, \oplus, \otimes, \bar{0}, \bar{1} \rangle$

- $\mathbb{K}$ is a set (e.g. $\mathbb{N}, \mathbb{R}, \{0, 1\}$)
- $\oplus$ and $\otimes$ are binary operators
- $\oplus$ is commutative and has identity $\bar{0}$
  $a \oplus b = b \oplus a$ and $\bar{0} \oplus a = a \oplus \bar{0} = a$
- $\otimes$ is associative and has identity $\bar{1}$
  $(a \otimes b) \otimes c = a \otimes (b \otimes c)$ and $\bar{1} \otimes a = a \otimes 1 = a$

Semirings

An algebraic structure $\mathcal{K} = \langle \mathbb{K}, \oplus, \otimes, \bar{0}, \bar{1} \rangle$

- $\mathbb{K}$ is a set (e.g. $\mathbb{N}, \mathbb{R}, \{0, 1\}$)
- $\oplus$ and $\otimes$ are binary operators
- $\oplus$ is commutative and has identity $\bar{0}$
  $a \oplus b = b \oplus a$ and $\bar{0} \oplus a = a \oplus \bar{0} = a$
- $\otimes$ is associative and has identity $\bar{1}$
  $(a \otimes b) \otimes c = a \otimes (b \otimes c)$ and $\bar{1} \otimes a = a \otimes 1 = a$
- $\otimes$ left distributes over $\oplus$
  $a \otimes (b \oplus c) = (a \otimes b) \oplus (a \otimes c)$

Semirings

An algebraic structure $\mathcal{K} = \langle \mathbb{K}, \oplus, \otimes, \bar{0}, \bar{1} \rangle$

- $\mathbb{K}$ is a set (e.g. $\mathbb{N}, \mathbb{R}, \{0, 1\}$)
- $\oplus$ and $\otimes$ are binary operators
- $\oplus$ is commutative and has identity $\bar{0}$
  $a \oplus b = b \oplus a$ and $\bar{0} \oplus a = a \oplus \bar{0} = a$
- $\otimes$ is associative and has identity $\bar{1}$
  $(a \otimes b) \otimes c = a \otimes (b \otimes c)$ and $\bar{1} \otimes a = a \otimes 1 = a$
- $\otimes$ left distributes over $\oplus$
  $a \otimes (b \oplus c) = (a \otimes b) \oplus (a \otimes c)$
- $\bar{0}$ is the $\otimes$-annihilator
  $\bar{0} \otimes a = a \otimes \bar{0} = \bar{0}$

Examples of semirings

Name	$\mathbb{K}$	$\oplus$	$\otimes$	$\bar{0}$	$\bar{1}$
Binary	$\{0, 1\}$	$\vee$	$\wedge$	0	1
Counting	$\mathbb{N}$	$+$	$\times$	0	1
Prob	$[0, 1] \subset \mathbb{R}$	$+$	$\times$	0	1
LogProb	$\mathbb{R} \cup \{-\infty\}$	$\oplus_{\log}$	$+$	$-\infty$	0
Viterbi	$\mathbb{R} \cup \{-\infty\}$	$\max$	$+$	$-\infty$	0

where $a \oplus_{\log} b = \log(\exp(a) + \exp(b))$

Linear models

$$f(\mathbf{d}) = \mathbf{w}^\top \mathbf{\Phi}(\mathbf{d})$$

where

- $\mathbf{w} \in \mathbb{R}^m$
- $\mathbf{\Phi}(\mathbf{d}) = \langle \Phi_1(\mathbf{d}), \ldots, \Phi_m(\mathbf{d}) \rangle$
- $\Phi_i(\mathbf{d}) \in \mathbb{R}$ is a feature function
- w_i is the relative contribution of the ith feature

Linear models and independence assumptions

$$f(\mathbf{d}) = \mathbf{w}^\top \mathbf{\Phi}(\mathbf{d}) \quad (1)$$

$$= \sum_{i=1}^{m} w_i \Phi_i(\mathbf{d}) \quad (2)$$

$$= \sum_{i=1}^{m} w_i \prod_{e \in \mathbf{d}} \phi_i(e) \quad (3)$$

$$= \prod_{e \in \mathbf{d}} \sum_{i=1}^{m} w_i \phi_i(e) \quad (4)$$

$$= \prod_{e \in \mathbf{d}} \mathbf{w}^\top \boldsymbol{\phi}(e) \quad (5)$$

Assumption

- $\Phi_i(\mathbf{d})$ factorises over edges
  $\phi_i(e)$ is a local feature function

Linear models and CFGs

Linear models can be expressed through hypergraphs using an appropriate semiring

Decision rules

Best translation (MAP)

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} \sum_{d \in \mathcal{D}_{\mathbf{y}}} f(\mathbf{d})$$

Decision rules

Best translation (MAP)

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} \sum_{d \in \mathcal{D}_{\mathbf{y}}} f(\mathbf{d})$$

Best derivation (Viterbi)

$$\mathbf{y}^* \approx \text{yield}\left\{\arg\max_{\mathbf{d}} f(\mathbf{d})\right\}$$

- less disambiguation power
- VITERBI semiring

Other decision rules?

Minimum Bayes risk (MBR)

$$\mathbf{y}^* = \arg\min_{\mathbf{y}'} \langle L(\mathbf{y}', \mathbf{y}) \rangle_{p(\mathbf{y})}$$

- requires the underlying model to have a probabilistic interpretation
- can be estimated through sampling

Other decision rules?

Minimum Bayes risk (MBR)

$$\mathbf{y}^* = \arg\min_{\mathbf{y}'} \langle L(\mathbf{y}', \mathbf{y}) \rangle_{p(\mathbf{y})}$$

- requires the underlying model to have a probabilistic interpretation
- can be estimated through sampling

Log-linear models

$$p(\mathbf{d}) = \frac{\exp(f(\mathbf{d}))}{\sum_{\mathbf{d}'} \exp(f(\mathbf{d}'))} \propto \exp\left(\sum_{e \in \mathbf{d}} \mathbf{w}^\top \boldsymbol{\phi}(e)\right) = \prod_{e \in \mathbf{d}} \exp(\mathbf{w}^\top \boldsymbol{\phi}(e))$$

LogProb semiring

Decoding

In SMT, decoding typically means the Viterbi approximation

$$\mathbf{d}^* = \arg\max_{\mathbf{d}} f(\mathbf{d})$$

Decoding

In SMT, decoding typically means the Viterbi approximation

$$\mathbf{d}^* = \arg\max_{\mathbf{d}} f(\mathbf{d})$$

If the statistical model $f(\mathbf{d})$ does not violate the independence assumptions of the model of translational equivalences

- steps in a derivation are weighted independently

Decoding

In SMT, decoding typically means the Viterbi approximation

$$\mathbf{d}^* = \arg\max_{\mathbf{d}} f(\mathbf{d})$$

If the statistical model $f(\mathbf{d})$ does not violate the independence assumptions of the model of translational equivalences

- steps in a derivation are weighted independently

there is a straightforward (tractable) decomposition of $f(\mathbf{d})$

- wFSA (phrase-based MT)
- wCFG (hierarchical MT)

Inside

The INSIDE recursion can be generalised to an arbitrary semiring

$$\beta(v) = \begin{cases} \bar{1} & \text{if } BS(v) = \emptyset \\ \displaystyle\bigoplus_{e \in BS(v)} w(e) \bigotimes_{u \in \text{tail}(e)} \beta(u) & \text{otherwise} \end{cases}$$

Inside

The Inside recursion can be generalised to an arbitrary semiring

$$\beta(v) = \begin{cases} \bar{1} & \text{if } BS(v) = \emptyset \\ \bigoplus_{e \in BS(v)} w(e) \bigotimes_{u \in \text{tail}(e)} \beta(u) & \text{otherwise} \end{cases}$$

- efficient bottom-up dynamic program $O(|G|)$
  $|G|$ is the size of the graphical representation of $f(\mathbf{d})$
  - a lattice (phrase-based MT)
  - a forest (hierarchical MT)

Inference

Viterbi derivation

1. start from the goal (root)
2. recursively rewrite every symbol v by solving

$$e = \arg\max_{e \in BS(v)} w(e) \bigotimes_{u \in \mathrm{tail}(e)} \beta(u)$$

Inference

Viterbi derivation

1. start from the goal (root)
2. recursively rewrite every symbol v by solving

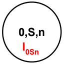
$$e = \arg\max_{e \in BS(v)} w(e) \bigotimes_{u \in \text{tail}(e)} \beta(u)$$

Sampling

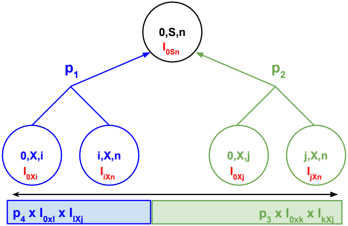
1. start from the goal (root)
2. recursively rewrite every symbol v by solving

$$e \sim p(e \in BS(v)|v) = \frac{w(e) \bigotimes_{u \in \text{tail}(e)} \beta(u)}{\beta(v)}$$

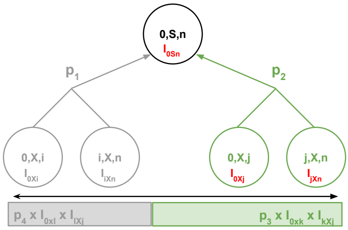
Viterbi



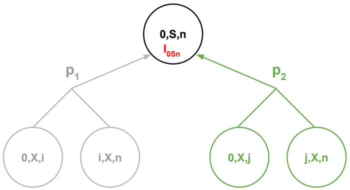
Viterbi


0,S,n
I0Sn
p1
p2
0,X,i
I0Xi
i,X,n
IiXn
0,X,j
I0Xj
j,X,n
IjXn
p4 x I0xl x IlXj
p3 x I0xk x IkXj

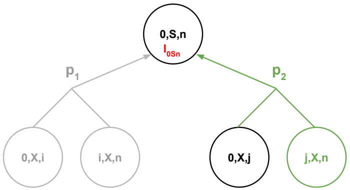
Viterbi


0,S,n
l0Sn
p1
p2
0,X,i
l0Xi
i,X,n
liXn
0,X,j
l0Xj
j,X,n
ljXn
p4 x l0xl x llXj
p3 x l0xk x lkXj

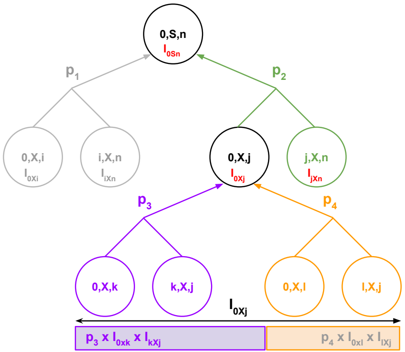
Viterbi



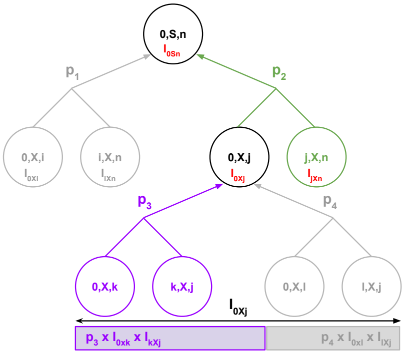
Viterbi



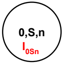
Viterbi


0,S,n
l0Sn
p1
p2
0,X,i
l0Xi
i,X,n
liXn
0,X,j
l0Xj
j,X,n
ljXn
p3
p4
0,X,k
k,X,j
0,X,l
l,X,j
l0Xj
p3 x l0xk x lkXj
p4 x l0xl x llXj

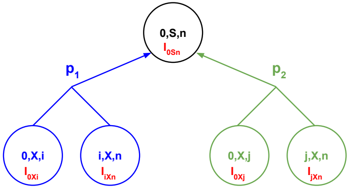
Viterbi


0,S,n
I0Sn
p1
p2
0,X,i
I0Xi
i,X,n
IiXn
0,X,j
I0Xj
j,X,n
IjXn
p3
p4
0,X,k
k,X,j
0,X,l
l,X,j
I0Xj
p3 x I0xk x IkXj
p4 x I0xl x IlXj

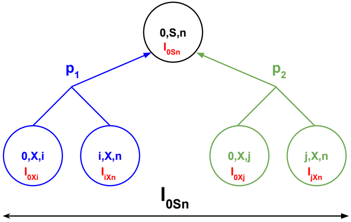
Sampling



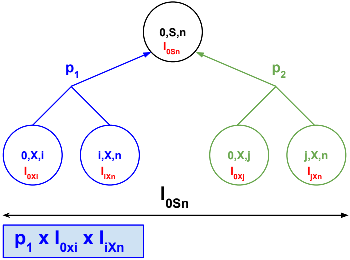
Sampling



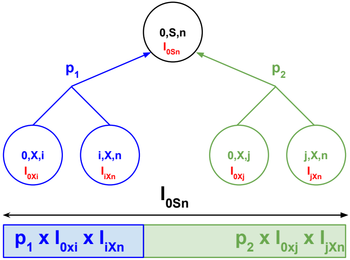
Sampling



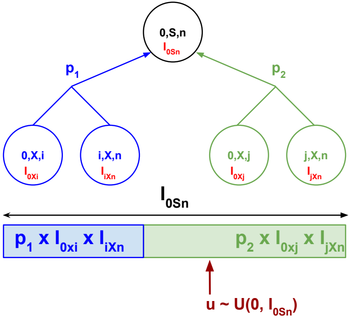
Sampling


0,S,n
I0Sn
p1
p2
0,X,i
I0Xi
i,X,n
IiXn
0,X,j
I0Xj
j,X,n
IjXn
I0Sn
p1 x I0xi x IiXn

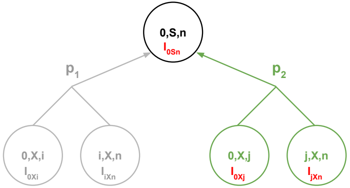
Sampling


0,S,n
I0Sn
p1
p2
0,X,i
I0Xi
i,X,n
IiXn
0,X,j
I0Xj
j,X,n
IjXn
I0Sn
p1 x I0xi x IiXn
p2 x I0xj x IjXn

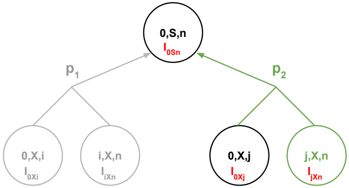
Sampling


0,S,n
I0Sn
p1
p2
0,X,i
I0Xi
i,X,n
IiXn
0,X,j
I0Xj
j,X,n
IjXn
I0Sn
p1 x I0xi x IiXn
p2 x I0xj x IjXn
u ~ U(0, I0Sn)

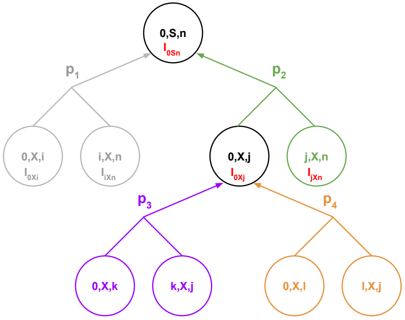
Sampling



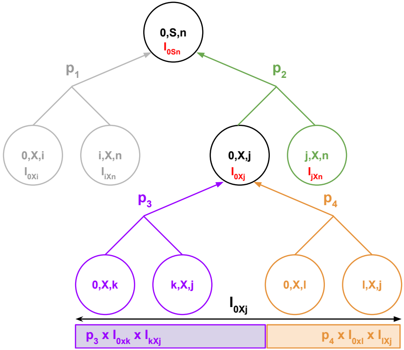
Sampling



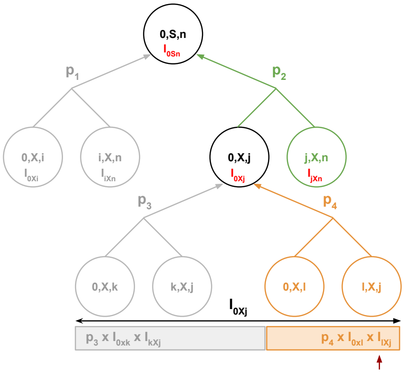
Sampling


0,S,n
I_{0Sn}
p_1
p_2
0,X,i
I_{0Xi}
i,X,n
I_{iXn}
0,X,j
I_{0Xj}
j,X,n
I_{jXn}
p_3
p_4
0,X,k
k,X,j
0,X,l
l,X,j

Sampling


0,S,n
I_{0Sn}
p_1
p_2
0,X,i
I_{0Xi}
i,X,n
I_{iXn}
0,X,j
I_{0Xj}
j,X,n
I_{jXn}
p_3
p_4
0,X,k
k,X,j
0,X,l
l,X,j
I_{0Xj}
$p_3 \times I_{0xk} \times I_{kXj}$
$p_4 \times I_{0xl} \times I_{lXj}$

Sampling


0,S,n
l0Sn
p1
p2
0,X,i
l0Xi
i,X,n
liXn
0,X,j
l0Xj
j,X,n
ljXn
p3
p4
0,X,k
k,X,j
0,X,l
l,X,j
l0Xj
p3 x l0xk x lkXj
p4 x l0xl x llXj

An example for hierarchical models

Model $f(\mathbf{d}) = \sum_i \varphi(e_i)$
where $\varphi(e_i)$ is a weighted combination of local features

Grammar

$X \rightarrow \langle \text{a}, \text{the} \rangle$
$X \rightarrow \langle \text{luz}, \text{light} \rangle$
$X \rightarrow \langle \text{apague } X_1, \text{switch } X_1 \text{ off} \rangle$
$X \rightarrow \langle X_1 \text{ por favor}, \text{please , } X_1 \rangle$
$X \rightarrow \langle X_1 X_2, X_1, X_2 \rangle$
$S \rightarrow \langle \vdash X_1 \dashv, \vdash X_1 \dashv \rangle$

Input: apague a luz por favor
Reference: please, switch the light off

Decoding with local features

Node	apague	a	luz	por	favor	Inside

Decoding with local features

NODE	apague	a	luz	por	favor	INSIDE
$X_{1,2}$		$e_1 : \frac{\text{a}}{\text{the}}$				

Decoding with local features

NODE	apague	a	luz	por	favor	INSIDE
$X_{1,2}$		$e_1 : \frac{\text{a}}{\text{the}}$				
$X_{2,3}$			$e_2 : \frac{\text{luz}}{\text{light}}$			

Decoding with local features

Node	apague	a	luz	por	favor	Inside
$X_{1,2}$		$e_1 : \frac{\text{a}}{\text{the}}$				
$X_{2,3}$			$e_2 : \frac{\text{luz}}{\text{light}}$			
$X_{1,3}$		$e_3 : \frac{X_{1,2}X_{2,3}}{X_{1,2}X_{2,3}}$				

Decoding with local features

Node	apague	a	luz	por	favor	Inside
$X_{1,2}$		$e_1 : \frac{\text{a}}{\text{the}}$				
$X_{2,3}$			$e_2 : \frac{\text{luz}}{\text{light}}$			
$X_{1,3}$		$e_3 : \frac{X_{1,2} X_{2,3}}{X_{1,2} X_{2,3}}$				
$X_{0,2}$	$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } X_{1,2} \text{ off}}$					

Decoding with local features

NODE	apague	a	luz	por	favor	INSIDE
$X_{1,2}$		$e_1 : \frac{\text{a}}{\text{the}}$				
$X_{2,3}$			$e_2 : \frac{\text{luz}}{\text{light}}$			
$X_{1,3}$		$e_3 : \frac{X_{1,2} X_{2,3}}{X_{1,2} X_{2,3}}$				
$X_{0,2}$	$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } X_{1,2} \text{off}}$					
$X_{2,5}$				$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } X_{2,3}}$		

Decoding with local features

NODE	apague	a	luz	por	favor	INSIDE
$X_{1,2}$		$e_1 : \frac{\text{a}}{\text{the}}$				
$X_{2,3}$			$e_2 : \frac{\text{luz}}{\text{light}}$			
$X_{1,3}$		$e_3 : \frac{X_{1,2} X_{2,3}}{X_{1,2} X_{2,3}}$				
$X_{0,2}$	$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } X_{1,2} \text{ off}}$					
$X_{2,5}$			$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } X_{2,3}}$			
$X_{0,3}$	$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } X_{1,3} \text{ off}}$					

Decoding with local features

<table>
<tr><th>Node</th><th>apague</th><th>a</th><th>luz</th><th>por</th><th>favor</th><th>Inside</th></tr>
<tr><td>$X_{1,2}$</td><td></td><td>$e_1 : \frac{\text{a}}{\text{the}}$</td><td></td><td></td><td></td><td></td></tr>
<tr><td>$X_{2,3}$</td><td></td><td></td><td>$e_2 : \frac{\text{luz}}{\text{light}}$</td><td></td><td></td><td></td></tr>
<tr><td>$X_{1,3}$</td><td></td><td colspan="2">$e_3 : \frac{X_{1,2} X_{2,3}}{X_{1,2} X_{2,3}}$</td><td></td><td></td><td></td></tr>
<tr><td>$X_{0,2}$</td><td colspan="2">$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } X_{1,2} \text{ off}}$</td><td></td><td></td><td></td><td></td></tr>
<tr><td>$X_{2,5}$</td><td></td><td></td><td colspan="3">$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } X_{2,3}}$</td><td></td></tr>
<tr><td rowspan="2">$X_{0,3}$</td><td colspan="3">$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } X_{1,3} \text{ off}}$</td><td></td><td></td><td rowspan="2"></td></tr>
<tr><td colspan="3">$e_7 : \frac{X_{0,2} X_{2,3}}{X_{0,2} X_{2,3}}$</td><td></td><td></td></tr>
<tr><td></td><td colspan="5"></td><td></td></tr>
<tr><td></td><td colspan="5"></td><td></td></tr>
<tr><td></td><td colspan="5"></td><td></td></tr>
</table>

Decoding with local features

Node	apague	a	luz	por	favor	Inside
$X_{1,2}$		$e_1 : \frac{\text{a}}{\text{the}}$				
$X_{2,3}$			$e_2 : \frac{\text{luz}}{\text{light}}$			
$X_{1,3}$		$e_3 : \frac{X_{1,2} X_{2,3}}{X_{1,2} X_{2,3}}$				
$X_{0,2}$	$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } X_{1,2} \text{ off}}$					
$X_{2,5}$			$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } X_{2,3}}$			
$X_{0,3}$	$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } X_{1,3} \text{ off}}$ $e_7 : \frac{X_{0,2} X_{2,3}}{X_{0,2} X_{2,3}}$					
$X_{1,5}$		$e_8 : \frac{X_{1,3} \text{ por favor}}{\text{please, } X_{1,3}}$				

Decoding with local features

<table>
<tr><th>Node</th><th>apague</th><th>a</th><th>luz</th><th>por</th><th>favor</th><th>Inside</th></tr>
<tr><td>$X_{1,2}$</td><td></td><td>$e_1 : \frac{\text{a}}{\text{the}}$</td><td></td><td></td><td></td><td></td></tr>
<tr><td>$X_{2,3}$</td><td></td><td></td><td>$e_2 : \frac{\text{luz}}{\text{light}}$</td><td></td><td></td><td></td></tr>
<tr><td>$X_{1,3}$</td><td></td><td colspan="2">$e_3 : \frac{X_{1,2} X_{2,3}}{X_{1,2} X_{2,3}}$</td><td></td><td></td><td></td></tr>
<tr><td>$X_{0,2}$</td><td colspan="2">$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } X_{1,2} \text{off}}$</td><td></td><td></td><td></td><td></td></tr>
<tr><td>$X_{2,5}$</td><td></td><td></td><td colspan="3">$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } X_{2,3}}$</td><td></td></tr>
<tr><td rowspan="2">$X_{0,3}$</td><td colspan="3">$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } X_{1,3} \text{off}}$</td><td></td><td></td><td rowspan="2"></td></tr>
<tr><td colspan="3">$e_7 : \frac{X_{0,2} X_{2,3}}{X_{0,2} X_{2,3}}$</td><td></td><td></td></tr>
<tr><td rowspan="2">$X_{1,5}$</td><td rowspan="2"></td><td colspan="4">$e_8 : \frac{X_{1,3} \text{ por favor}}{\text{please, } X_{1,3}}$</td><td rowspan="2"></td></tr>
<tr><td colspan="4">$e_9 : \frac{X_{1,2} X_{2,5}}{X_{1,2} X_{2,5}}$</td></tr>
<tr><td></td><td colspan="5"></td><td></td></tr>
<tr><td></td><td colspan="5"></td><td></td></tr>
</table>

Decoding with local features

<table>
<tr><th>NODE</th><th>apague</th><th>a</th><th>luz</th><th>por</th><th>favor</th><th>INSIDE</th></tr>
<tr><td>$X_{1,2}$</td><td></td><td>$e_1 : \frac{\text{a}}{\text{the}}$</td><td></td><td></td><td></td><td></td></tr>
<tr><td>$X_{2,3}$</td><td></td><td></td><td>$e_2 : \frac{\text{luz}}{\text{light}}$</td><td></td><td></td><td></td></tr>
<tr><td>$X_{1,3}$</td><td></td><td colspan="2">$e_3 : \frac{X_{1,2} X_{2,3}}{X_{1,2} X_{2,3}}$</td><td></td><td></td><td></td></tr>
<tr><td>$X_{0,2}$</td><td colspan="2">$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } X_{1,2} \text{ off}}$</td><td></td><td></td><td></td><td></td></tr>
<tr><td>$X_{2,5}$</td><td></td><td></td><td colspan="3">$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } X_{2,3}}$</td><td></td></tr>
<tr><td rowspan="2">$X_{0,3}$</td><td colspan="3">$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } X_{1,3} \text{ off}}$</td><td></td><td></td><td rowspan="2"></td></tr>
<tr><td colspan="3">$e_7 : \frac{X_{0,2} X_{2,3}}{X_{0,2} X_{2,3}}$</td><td></td><td></td></tr>
<tr><td rowspan="2">$X_{1,5}$</td><td rowspan="2"></td><td colspan="4">$e_8 : \frac{X_{1,3} \text{ por favor}}{\text{please, } X_{1,3}}$</td><td rowspan="2"></td></tr>
<tr><td colspan="4">$e_9 : \frac{X_{1,2} X_{2,5}}{X_{1,2} X_{2,5}}$</td></tr>
<tr><td>$X_{0,5}$</td><td colspan="5">$e_{10} : \frac{X_{0,2} X_{2,5}}{X_{0,2} X_{2,5}}$</td><td></td></tr>
<tr><td></td><td colspan="5"></td><td></td></tr>
</table>

Decoding with local features

Node	apague	a	luz	por	favor	Inside
$X_{1,2}$		$e_1 : \frac{\text{a}}{\text{the}}$				
$X_{2,3}$			$e_2 : \frac{\text{luz}}{\text{light}}$			
$X_{1,3}$		$e_3 : \frac{X_{1,2} X_{2,3}}{X_{1,2} X_{2,3}}$				
$X_{0,2}$	$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } X_{1,2} \text{ off}}$					
$X_{2,5}$			$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } X_{2,3}}$			
$X_{0,3}$	$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } X_{1,3} \text{ off}}$ $e_7 : \frac{X_{0,2} X_{2,3}}{X_{0,2} X_{2,3}}$					
$X_{1,5}$		$e_8 : \frac{X_{1,3} \text{ por favor}}{\text{please, } X_{1,3}}$ $e_9 : \frac{X_{1,2} X_{2,5}}{X_{1,2} X_{2,5}}$				
$X_{0,5}$	$e_{10} : \frac{X_{0,2} X_{2,5}}{X_{0,2} X_{2,5}}$ $e_{11} : \frac{\text{apague } X_{1,5}}{\text{switch } X_{1,5} \text{ off}}$					

Decoding with local features

Node	apague	a	luz	por	favor	Inside
$X_{1,2}$		$e_1 : \frac{\text{a}}{\text{the}}$				
$X_{2,3}$			$e_2 : \frac{\text{luz}}{\text{light}}$			
$X_{1,3}$		$e_3 : \frac{X_{1,2} X_{2,3}}{X_{1,2} X_{2,3}}$				
$X_{0,2}$	$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } X_{1,2} \text{off}}$					
$X_{2,5}$			$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } X_{2,3}}$			
$X_{0,3}$	$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } X_{1,3} \text{off}}$ $e_7 : \frac{X_{0,2} X_{2,3}}{X_{0,2} X_{2,3}}$					
$X_{1,5}$		$e_8 : \frac{X_{1,3} \text{ por favor}}{\text{please, } X_{1,3}}$ $e_9 : \frac{X_{1,2} X_{2,5}}{X_{1,2} X_{2,5}}$				
$X_{0,5}$	$e_{10} : \frac{X_{0,2} X_{2,5}}{X_{0,2} X_{2,5}}$ $e_{11} : \frac{\text{apague } X_{1,5}}{\text{switch } X_{1,5} \text{ off}}$ $e_{12} : \frac{X_{0,3} \text{ por favor}}{\text{please, } X_{0,3}}$					

Decoding with local features

NODE	apague	a	luz	por	favor	INSIDE
$X_{1,2}$		$e_1 : \frac{\text{a}}{\text{the}}$				
$X_{2,3}$			$e_2 : \frac{\text{luz}}{\text{light}}$			
$X_{1,3}$		$e_3 : \frac{X_{1,2} X_{2,3}}{X_{1,2} X_{2,3}}$				
$X_{0,2}$	$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } X_{1,2} \text{ off}}$					
$X_{2,5}$			$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } X_{2,3}}$			
$X_{0,3}$	$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } X_{1,3} \text{ off}}$					
	$e_7 : \frac{X_{0,2} X_{2,3}}{X_{0,2} X_{2,3}}$					
$X_{1,5}$		$e_8 : \frac{X_{1,3} \text{ por favor}}{\text{please, } X_{1,3}}$				
		$e_9 : \frac{X_{1,2} X_{2,5}}{X_{1,2} X_{2,5}}$				
$X_{0,5}$	$e_{10} : \frac{X_{0,2} X_{2,5}}{X_{0,2} X_{2,5}}$					
	$e_{11} : \frac{\text{apague } X_{1,5}}{\text{switch } X_{1,5} \text{ off}}$					
	$e_{12} : \frac{X_{0,3} \text{ por favor}}{\text{please, } X_{0,3}}$					
$S_{0,5}$	$e_{13} : \frac{\vdash X_{0,5} \dashv}{\vdash X_{0,5} \dashv}$					

Decoding with local features

NODE	apague a luz por favor	INSIDE
$X_{1,2}$	$e_1 : \frac{\text{a}}{\text{the}}$	$w(e_1)$
$X_{2,3}$	$e_2 : \frac{\text{luz}}{\text{light}}$	
$X_{1,3}$	$e_3 : \frac{X_{1,2} X_{2,3}}{X_{1,2} X_{2,3}}$	
$X_{0,2}$	$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } X_{1,2} \text{ off}}$	
$X_{2,5}$	$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } X_{2,3}}$	
$X_{0,3}$	$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } X_{1,3} \text{ off}}$ $e_7 : \frac{X_{0,2} X_{2,3}}{X_{0,2} X_{2,3}}$	
$X_{1,5}$	$e_8 : \frac{X_{1,3} \text{ por favor}}{\text{please, } X_{1,3}}$ $e_9 : \frac{X_{1,2} X_{2,5}}{X_{1,2} X_{2,5}}$	
$X_{0,5}$	$e_{10} : \frac{X_{0,2} X_{2,5}}{X_{0,2} X_{2,5}}$ $e_{11} : \frac{\text{apague } X_{1,5}}{\text{switch } X_{1,5} \text{ off}}$ $e_{12} : \frac{X_{0,3} \text{ por favor}}{\text{please, } X_{0,3}}$	
$S_{0,5}$	$e_{13} : \frac{\vdash X_{0,5} \dashv}{\vdash X_{0,5} \dashv}$	

Decoding with local features

Node	apague	a	luz	por	favor	Inside
$X_{1,2}$		$e_1 : \frac{\text{a}}{\text{the}}$				$w(e_1)$
$X_{2,3}$			$e_2 : \frac{\text{luz}}{\text{light}}$			$w(e_2)$
$X_{1,3}$		$e_3 : \frac{X_{1,2} X_{2,3}}{X_{1,2} X_{2,3}}$				
$X_{0,2}$	$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } X_{1,2} \text{ off}}$					
$X_{2,5}$			$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } X_{2,3}}$			
$X_{0,3}$	$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } X_{1,3} \text{ off}}$					
	$e_7 : \frac{X_{0,2} X_{2,3}}{X_{0,2} X_{2,3}}$					
$X_{1,5}$		$e_8 : \frac{X_{1,3} \text{ por favor}}{\text{please, } X_{1,3}}$				
		$e_9 : \frac{X_{1,2} X_{2,5}}{X_{1,2} X_{2,5}}$				
$X_{0,5}$	$e_{10} : \frac{X_{0,2} X_{2,5}}{X_{0,2} X_{2,5}}$					
	$e_{11} : \frac{\text{apague } X_{1,5}}{\text{switch } X_{1,5} \text{ off}}$					
	$e_{12} : \frac{X_{0,3} \text{ por favor}}{\text{please, } X_{0,3}}$					
$S_{0,5}$	$e_{13} : \frac{\vdash X_{0,5} \dashv}{\vdash X_{0,5} \dashv}$					

Decoding with local features

<table>
<tr><th>Node</th><th>apague</th><th>a</th><th>luz</th><th>por</th><th>favor</th><th>Inside</th></tr>
<tr><td>$X_{1,2}$</td><td></td><td>$e_1 : \frac{\text{a}}{\text{the}}$</td><td></td><td></td><td></td><td>$w(e_1)$</td></tr>
<tr><td>$X_{2,3}$</td><td></td><td></td><td>$e_2 : \frac{\text{luz}}{\text{light}}$</td><td></td><td></td><td>$w(e_2)$</td></tr>
<tr><td>$X_{1,3}$</td><td></td><td colspan="2">$e_3 : \frac{X_{1,2} X_{2,3}}{X_{1,2} X_{2,3}}$</td><td></td><td></td><td>$w(e_3)\beta(X_{1,2})\beta(X_{2,3})$</td></tr>
<tr><td>$X_{0,2}$</td><td colspan="2">$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } X_{1,2} \text{ off}}$</td><td></td><td></td><td></td><td></td></tr>
<tr><td>$X_{2,5}$</td><td></td><td></td><td colspan="3">$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } X_{2,3}}$</td><td></td></tr>
<tr><td rowspan="2">$X_{0,3}$</td><td colspan="3">$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } X_{1,3} \text{ off}}$</td><td></td><td></td><td rowspan="2"></td></tr>
<tr><td colspan="3">$e_7 : \frac{X_{0,2} X_{2,3}}{X_{0,2} X_{2,3}}$</td><td></td><td></td></tr>
<tr><td rowspan="2">$X_{1,5}$</td><td></td><td colspan="4">$e_8 : \frac{X_{1,3} \text{ por favor}}{\text{please, } X_{1,3}}$</td><td rowspan="2"></td></tr>
<tr><td></td><td colspan="4">$e_9 : \frac{X_{1,2} X_{2,5}}{X_{1,2} X_{2,5}}$</td></tr>
<tr><td rowspan="3">$X_{0,5}$</td><td colspan="5">$e_{10} : \frac{X_{0,2} X_{2,5}}{X_{0,2} X_{2,5}}$</td><td rowspan="3"></td></tr>
<tr><td colspan="5">$e_{11} : \frac{\text{apague } X_{1,5}}{\text{switch } X_{1,5} \text{ off}}$</td></tr>
<tr><td colspan="5">$e_{12} : \frac{X_{0,3} \text{ por favor}}{\text{please, } X_{0,3}}$</td></tr>
<tr><td>$S_{0,5}$</td><td colspan="5">$e_{13} : \frac{\vdash X_{0,5} \dashv}{\vdash X_{0,5} \dashv}$</td><td></td></tr>
</table>

Decoding with local features

Node	apague	a	luz	por	favor	Inside
$X_{1,2}$		$e_1 : \frac{\text{a}}{\text{the}}$				$w(e_1)$
$X_{2,3}$			$e_2 : \frac{\text{luz}}{\text{light}}$			$w(e_2)$
$X_{1,3}$		$e_3 : \frac{X_{1,2} X_{2,3}}{X_{1,2} X_{2,3}}$				$w(e_3)\beta(X_{1,2})\beta(X_{2,3})$
$X_{0,2}$	$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } X_{1,2} \text{off}}$					$w(e_4)\beta(X_{1,2})$
$X_{2,5}$			$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } X_{2,3}}$			
$X_{0,3}$	$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } X_{1,3} \text{off}}$					
	$e_7 : \frac{X_{0,2} X_{2,3}}{X_{0,2} X_{2,3}}$					
$X_{1,5}$		$e_8 : \frac{X_{1,3} \text{ por favor}}{\text{please, } X_{1,3}}$				
		$e_9 : \frac{X_{1,2} X_{2,5}}{X_{1,2} X_{2,5}}$				
$X_{0,5}$	$e_{10} : \frac{X_{0,2} X_{2,5}}{X_{0,2} X_{2,5}}$					
	$e_{11} : \frac{\text{apague } X_{1,5}}{\text{switch } X_{1,5} \text{ off}}$					
	$e_{12} : \frac{X_{0,3} \text{ por favor}}{\text{please, } X_{0,3}}$					
$S_{0,5}$	$e_{13} : \frac{\vdash X_{0,5} \dashv}{\vdash X_{0,5} \dashv}$					

Decoding with local features

<table>
<tr><th>Node</th><th>apague</th><th>a</th><th>luz</th><th>por</th><th>favor</th><th>Inside</th></tr>
<tr><td>$X_{1,2}$</td><td></td><td>$e_1 : \frac{\text{a}}{\text{the}}$</td><td></td><td></td><td></td><td>$w(e_1)$</td></tr>
<tr><td>$X_{2,3}$</td><td></td><td></td><td>$e_2 : \frac{\text{luz}}{\text{light}}$</td><td></td><td></td><td>$w(e_2)$</td></tr>
<tr><td>$X_{1,3}$</td><td></td><td colspan="2">$e_3 : \frac{X_{1,2}X_{2,3}}{X_{1,2}X_{2,3}}$</td><td></td><td></td><td>$w(e_3)\beta(X_{1,2})\beta(X_{2,3})$</td></tr>
<tr><td>$X_{0,2}$</td><td colspan="2">$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } X_{1,2}\text{off}}$</td><td></td><td></td><td></td><td>$w(e_4)\beta(X_{1,2})$</td></tr>
<tr><td>$X_{2,5}$</td><td></td><td></td><td colspan="3">$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } X_{2,3}}$</td><td>$w(e_5)\beta(X_{2,3})$</td></tr>
<tr><td rowspan="2">$X_{0,3}$</td><td colspan="3">$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } X_{1,3}\text{off}}$</td><td></td><td></td><td rowspan="2"></td></tr>
<tr><td colspan="3">$e_7 : \frac{X_{0,2}X_{2,3}}{X_{0,2}X_{2,3}}$</td><td></td><td></td></tr>
<tr><td rowspan="2">$X_{1,5}$</td><td></td><td colspan="4">$e_8 : \frac{X_{1,3} \text{ por favor}}{\text{please, } X_{1,3}}$</td><td rowspan="2"></td></tr>
<tr><td></td><td colspan="4">$e_9 : \frac{X_{1,2}X_{2,5}}{X_{1,2}X_{2,5}}$</td></tr>
<tr><td rowspan="3">$X_{0,5}$</td><td colspan="5">$e_{10} : \frac{X_{0,2}X_{2,5}}{X_{0,2}X_{2,5}}$</td><td rowspan="3"></td></tr>
<tr><td colspan="5">$e_{11} : \frac{\text{apague } X_{1,5}}{\text{switch } X_{1,5} \text{ off}}$</td></tr>
<tr><td colspan="5">$e_{12} : \frac{X_{0,3} \text{ por favor}}{\text{please, } X_{0,3}}$</td></tr>
<tr><td>$S_{0,5}$</td><td colspan="5">$e_{13} : \frac{\vdash X_{0,5} \dashv}{\vdash X_{0,5} \dashv}$</td><td></td></tr>
</table>

Decoding with local features

<table>
<tr><th>Node</th><th>apague</th><th>a</th><th>luz</th><th>por</th><th>favor</th><th>Inside</th></tr>
<tr><td>$X_{1,2}$</td><td></td><td>$e_1 : \frac{\text{a}}{\text{the}}$</td><td></td><td></td><td></td><td>$w(e_1)$</td></tr>
<tr><td>$X_{2,3}$</td><td></td><td></td><td>$e_2 : \frac{\text{luz}}{\text{light}}$</td><td></td><td></td><td>$w(e_2)$</td></tr>
<tr><td>$X_{1,3}$</td><td></td><td colspan="2">$e_3 : \frac{X_{1,2} X_{2,3}}{X_{1,2} X_{2,3}}$</td><td></td><td></td><td>$w(e_3)\beta(X_{1,2})\beta(X_{2,3})$</td></tr>
<tr><td>$X_{0,2}$</td><td colspan="2">$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } X_{1,2} \text{ off}}$</td><td></td><td></td><td></td><td>$w(e_4)\beta(X_{1,2})$</td></tr>
<tr><td>$X_{2,5}$</td><td></td><td></td><td colspan="3">$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } X_{2,3}}$</td><td>$w(e_5)\beta(X_{2,3})$</td></tr>
<tr><td rowspan="2">$X_{0,3}$</td><td colspan="3">$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } X_{1,3} \text{ off}}$</td><td></td><td></td><td>$w(e_6)\beta(X_{1,3}) \oplus$</td></tr>
<tr><td colspan="3">$e_7 : \frac{X_{0,2} X_{2,3}}{X_{0,2} X_{2,3}}$</td><td></td><td></td><td>$w(e_7)\beta(X_{0,2})\beta(X_{2,3})$</td></tr>
<tr><td rowspan="2">$X_{1,5}$</td><td></td><td colspan="4">$e_8 : \frac{X_{1,3} \text{ por favor}}{\text{please, } X_{1,3}}$</td><td rowspan="2"></td></tr>
<tr><td></td><td colspan="4">$e_9 : \frac{X_{1,2} X_{2,5}}{X_{1,2} X_{2,5}}$</td></tr>
<tr><td rowspan="3">$X_{0,5}$</td><td colspan="5">$e_{10} : \frac{X_{0,2} X_{2,5}}{X_{0,2} X_{2,5}}$</td><td rowspan="3"></td></tr>
<tr><td colspan="5">$e_{11} : \frac{\text{apague } X_{1,5}}{\text{switch } X_{1,5} \text{ off}}$</td></tr>
<tr><td colspan="5">$e_{12} : \frac{X_{0,3} \text{ por favor}}{\text{please, } X_{0,3}}$</td></tr>
<tr><td>$S_{0,5}$</td><td colspan="5">$e_{13} : \frac{\vdash X_{0,5} \dashv}{\vdash X_{0,5} \dashv}$</td><td></td></tr>
</table>

Decoding with local features

Node	apague	a	luz	por	favor	Inside
$X_{1,2}$		$e_1 : \frac{\text{a}}{\text{the}}$				$w(e_1)$
$X_{2,3}$			$e_2 : \frac{\text{luz}}{\text{light}}$			$w(e_2)$
$X_{1,3}$		$e_3 : \frac{X_{1,2} X_{2,3}}{X_{1,2} X_{2,3}}$				$w(e_3)\beta(X_{1,2})\beta(X_{2,3})$
$X_{0,2}$	$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } X_{1,2} \text{off}}$					$w(e_4)\beta(X_{1,2})$
$X_{2,5}$			$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } X_{2,3}}$			$w(e_5)\beta(X_{2,3})$
$X_{0,3}$	$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } X_{1,3} \text{off}}$					$w(e_6)\beta(X_{1,3}) \oplus$
	$e_7 : \frac{X_{0,2} X_{2,3}}{X_{0,2} X_{2,3}}$					$w(e_7)\beta(X_{0,2})\beta(X_{2,3})$
$X_{1,5}$		$e_8 : \frac{X_{1,3} \text{ por favor}}{\text{please, } X_{1,3}}$				$w(e_8)\beta(X_{1,3}) \oplus$
		$e_9 : \frac{X_{1,2} X_{2,5}}{X_{1,2} X_{2,5}}$				$w(e_9)\beta(X_{1,2})\beta(X_{2,5})$
$X_{0,5}$	$e_{10} : \frac{X_{0,2} X_{2,5}}{X_{0,2} X_{2,5}}$					
	$e_{11} : \frac{\text{apague } X_{1,5}}{\text{switch } X_{1,5} \text{ off}}$					
	$e_{12} : \frac{X_{0,3} \text{ por favor}}{\text{please, } X_{0,3}}$					
$S_{0,5}$	$e_{13} : \frac{\vdash X_{0,5} \dashv}{\vdash X_{0,5} \dashv}$					

Decoding with local features

Node	apague a luz por favor	Inside
$X_{1,2}$	$e_1 : \frac{\text{a}}{\text{the}}$	$w(e_1)$
$X_{2,3}$	$e_2 : \frac{\text{luz}}{\text{light}}$	$w(e_2)$
$X_{1,3}$	$e_3 : \frac{X_{1,2} X_{2,3}}{X_{1,2} X_{2,3}}$	$w(e_3)\beta(X_{1,2})\beta(X_{2,3})$
$X_{0,2}$	$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } X_{1,2} \text{off}}$	$w(e_4)\beta(X_{1,2})$
$X_{2,5}$	$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } X_{2,3}}$	$w(e_5)\beta(X_{2,3})$
$X_{0,3}$	$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } X_{1,3} \text{off}}$ $e_7 : \frac{X_{0,2} X_{2,3}}{X_{0,2} X_{2,3}}$	$w(e_6)\beta(X_{1,3}) \oplus$ $w(e_7)\beta(X_{0,2})\beta(X_{2,3})$
$X_{1,5}$	$e_8 : \frac{X_{1,3} \text{ por favor}}{\text{please, } X_{1,3}}$ $e_9 : \frac{X_{1,2} X_{2,5}}{X_{1,2} X_{2,5}}$	$w(e_8)\beta(X_{1,3}) \oplus$ $w(e_9)\beta(X_{1,2})\beta(X_{2,5})$
$X_{0,5}$	$e_{10} : \frac{X_{0,2} X_{2,5}}{X_{0,2} X_{2,5}}$ $e_{11} : \frac{\text{apague } X_{1,5}}{\text{switch } X_{1,5} \text{ off}}$ $e_{12} : \frac{X_{0,3} \text{ por favor}}{\text{please, } X_{0,3}}$	$w(e_{10})\beta(X_{0,2})\beta(X_{2,5}) \oplus$ $w(e_{11})\beta(X_{1,5}) \oplus$ $w(e_{12})\beta(X_{0,3})$
$S_{0,5}$	$e_{13} : \frac{\vdash X_{0,5} \dashv}{\vdash X_{0,5} \dashv}$	

Decoding with local features

<table>
<tr><th>NODE</th><th>apague</th><th>a</th><th>luz</th><th>por</th><th>favor</th><th>INSIDE</th></tr>
<tr><td>$X_{1,2}$</td><td></td><td>$e_1 : \frac{\text{a}}{\text{the}}$</td><td></td><td></td><td></td><td>$w(e_1)$</td></tr>
<tr><td>$X_{2,3}$</td><td></td><td></td><td>$e_2 : \frac{\text{luz}}{\text{light}}$</td><td></td><td></td><td>$w(e_2)$</td></tr>
<tr><td>$X_{1,3}$</td><td></td><td colspan="2">$e_3 : \frac{X_{1,2} X_{2,3}}{X_{1,2} X_{2,3}}$</td><td></td><td></td><td>$w(e_3)\beta(X_{1,2})\beta(X_{2,3})$</td></tr>
<tr><td>$X_{0,2}$</td><td colspan="2">$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } X_{1,2} \text{off}}$</td><td></td><td></td><td></td><td>$w(e_4)\beta(X_{1,2})$</td></tr>
<tr><td>$X_{2,5}$</td><td></td><td></td><td colspan="3">$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } X_{2,3}}$</td><td>$w(e_5)\beta(X_{2,3})$</td></tr>
<tr><td rowspan="2">$X_{0,3}$</td><td colspan="3">$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } X_{1,3} \text{off}}$</td><td></td><td></td><td>$w(e_6)\beta(X_{1,3})\oplus$</td></tr>
<tr><td colspan="3">$e_7 : \frac{X_{0,2} X_{2,3}}{X_{0,2} X_{2,3}}$</td><td></td><td></td><td>$w(e_7)\beta(X_{0,2})\beta(X_{2,3})$</td></tr>
<tr><td rowspan="2">$X_{1,5}$</td><td></td><td colspan="4">$e_8 : \frac{X_{1,3} \text{ por favor}}{\text{please, } X_{1,3}}$</td><td>$w(e_8)\beta(X_{1,3})\oplus$</td></tr>
<tr><td></td><td colspan="4">$e_9 : \frac{X_{1,2} X_{2,5}}{X_{1,2} X_{2,5}}$</td><td>$w(e_9)\beta(X_{1,2})\beta(X_{2,5})$</td></tr>
<tr><td rowspan="3">$X_{0,5}$</td><td colspan="5">$e_{10} : \frac{X_{0,2} X_{2,5}}{X_{0,2} X_{2,5}}$</td><td>$w(e_{10})\beta(X_{0,2})\beta(X_{2,5})\oplus$</td></tr>
<tr><td colspan="5">$e_{11} : \frac{\text{apague } X_{1,5}}{\text{switch } X_{1,5} \text{ off}}$</td><td>$w(e_{11})\beta(X_{1,5})\oplus$</td></tr>
<tr><td colspan="5">$e_{12} : \frac{X_{0,3} \text{ por favor}}{\text{please, } X_{0,3}}$</td><td>$w(e_{12})\beta(X_{0,3})$</td></tr>
<tr><td>$S_{0,5}$</td><td colspan="5">$e_{13} : \frac{\vdash X_{0,5} \dashv}{\vdash X_{0,5} \dashv}$</td><td>$w(e_{13})\beta(X_{0,5})$</td></tr>
</table>

The problem

Most interesting models employ nonlocal features!

- reordering model: previously translated span
- language model: generated strings

The problem

Most interesting models employ nonlocal features!

- reordering model: previously translated span
- language model: generated strings

Example

$$f(\mathbf{d}) = \psi(\text{yield}(\mathbf{d})) + \sum_i \varphi(e_i)$$

where $\psi(\mathbf{y}) = w_\psi \log p_{\text{LM}}(\mathbf{y})$
and $p_{\text{LM}}(\mathbf{y}) = \prod_i p(y_i | y_{i-n+1}^{i-1})$ is an n-gram LM

The problem

Most interesting models employ nonlocal features!

- reordering model: previously translated span
- language model: generated strings

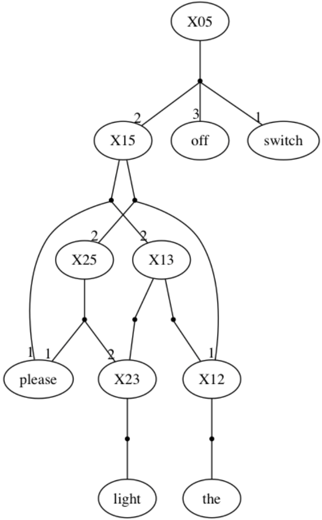
Example

$$f(\mathbf{d}) = \psi(\text{yield}(\mathbf{d})) + \sum_i \varphi(e_i)$$

where $\psi(\mathbf{y}) = w_\psi \log p_{\text{LM}}(\mathbf{y})$
and $p_{\text{LM}}(\mathbf{y}) = \prod_i p(y_i | y_{i-n+1}^{i-1})$ is an n-gram LM

- p_{LM} violates independence assumptions

Illustration of the problem



How do we score the top edge?

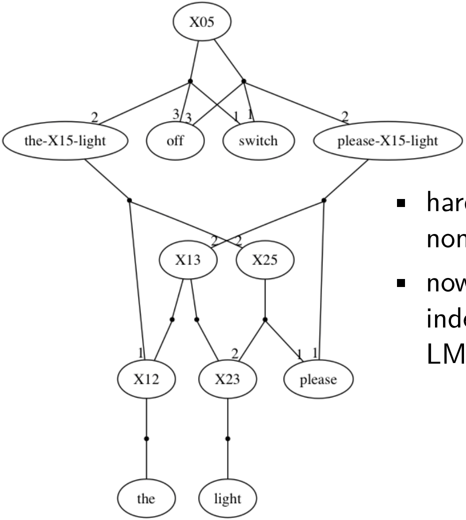
- [switch [please [[the][light]]] off]
- [switch [[the][please [light]]] off]

The solution

"Hard-code" structural dependencies

- disambiguate nodes w.r.t. the context they offer to feature functions
- intuition: we will be "splitting" nodes
- more intuition: nodes must memorise how to complete boundary n-grams

The intuition



- hard-code dependencies through nonterminals
- now we can weight edges independently even with a bigram LM

An example for hierarchical models

Model $f(\mathbf{d}) = \psi(\text{yield}(\mathbf{d})) \sum_i \varphi(e_i)$

where $\varphi(e_i)$ is a weighted combination of local features

$\psi(\text{yield}(\mathbf{d}))$ contains a 3-gram LM

i.e, $p_{\text{LM}_3}(\mathbf{y}) = \prod_i p(y_i | y_{i-2} y_{i-1})$

Grammar

$X \rightarrow \langle \text{a}, \text{the} \rangle$
$X \rightarrow \langle \text{luz}, \text{light} \rangle$
$X \rightarrow \langle \text{apague } X_1, \text{switch } X_1 \text{ off} \rangle$
$X \rightarrow \langle X_1 \text{ por favor}, \text{please , } X_1 \rangle$
$X \rightarrow \langle X_1 X_2, X_1, X_2 \rangle$
$S \rightarrow \langle \vdash X_1 \dashv, \vdash X_1 \dashv \rangle$

Input: apague a luz por favor

Reference: please, switch the light off

	apague a luz por favor	Left	Right	Node
				1
				2
				3
				4
				5
				6
				7
				8
				9
				10
				11
				12
				13
				14
				15
				16
				17
				18

	apague	a	luz	por	favor	LEFT	RIGHT	NODE
$X_{1,2}$		$e_1 : \frac{\text{a}}{\text{the}}$				the	the	1
								2
								3
								4
								5
								6
								7
								8
								9
								10
								11
								12
								13
								14
								15
								16
								17
								18

	apague	a	luz	por	favor	Left	Right	Node
$X_{1,2}$		$e_1 : \frac{\text{a}}{\text{the}}$				the	the	1
$X_{2,3}$			$e_2 : \frac{\text{luz}}{\text{light}}$			light	light	2
								3
								4
								5
								6
								7
								8
								9
								10
								11
								12
								13
								14
								15
								16
								17
								18

	apague	a	luz	por	favor	Left	Right	Node
$X_{1,2}$		$e_1 : \frac{\text{a}}{\text{the}}$				the	the	1
$X_{2,3}$			$e_2 : \frac{\text{luz}}{\text{light}}$			light	light	2
$X_{1,3}$		$e_3 : \frac{X_{1,2} X_{2,3}}{(1)\ \ (2)}$				the light	the light	3
								4
								5
								6
								7
								8
								9
								10
								11
								12
								13
								14
								15
								16
								17
								18

	apague	a	luz	por	favor	LEFT	RIGHT	NODE
$X_{1,2}$		$e_1 : \frac{\text{a}}{\text{the}}$				the	the	1
$X_{2,3}$			$e_2 : \frac{\text{luz}}{\text{light}}$			light	light	2
$X_{1,3}$		$e_3 : \frac{X_{1,2}\,X_{2,3}}{(1)\;(2)}$				the light	the light	3
$X_{0,2}$	$e_4 : \frac{\text{apague}\ X_{1,2}}{\text{switch}\ (1)\ \text{off}}$					switch the	the off	4
								5
								6
								7
								8
								9
								10
								11
								12
								13
								14
								15
								16
								17
								18

	apague	a	luz	por	favor	Left	Right	Node
$X_{1,2}$		$e_1 : \frac{\text{a}}{\text{the}}$				the	the	1
$X_{2,3}$			$e_2 : \frac{\text{luz}}{\text{light}}$			light	light	2
$X_{1,3}$		$e_3 : \frac{X_{1,2}\,X_{2,3}}{(1)\ \ (2)}$				the light	the light	3
$X_{0,2}$	$e_4 : \frac{\text{apague}\ X_{1,2}}{\text{switch}\ (1)\ \text{off}}$					switch the	the off	4
$X_{2,5}$			$e_5 : \frac{X_{2,3}\ \text{por favor}}{\text{please,}\ (2)}$			please ,	, light	5
								6
								7
								8
								9
								10
								11
								12
								13
								14
								15
								16
								17
								18

	apague	a	luz	por	favor	LEFT	RIGHT	NODE
$X_{1,2}$		$e_1 : \frac{\text{a}}{\text{the}}$				the	the	1
$X_{2,3}$			$e_2 : \frac{\text{luz}}{\text{light}}$			light	light	2
$X_{1,3}$		$e_3 : \frac{X_{1,2}\,X_{2,3}}{(1)\;(2)}$				the light	the light	3
$X_{0,2}$	$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } (1) \text{ off}}$					switch the	the off	4
$X_{2,5}$			$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } (2)}$			please ,	, light	5
$X_{0,3}$	$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } (3) \text{ off}}$					switch the	light off	6
								7
								8
								9
								10
								11
								12
								13
								14
								15
								16
								17
								18

	apague	a	luz	por	favor	LEFT	RIGHT	NODE
$X_{1,2}$		$e_1 : \frac{\text{a}}{\text{the}}$				the	the	1
$X_{2,3}$			$e_2 : \frac{\text{luz}}{\text{light}}$			light	light	2
$X_{1,3}$		$e_3 : \frac{X_{1,2} X_{2,3}}{(1)\ (2)}$				the light	the light	3
$X_{0,2}$	$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } (1) \text{ off}}$					switch the	the off	4
$X_{2,5}$			$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } (2)}$			please ,	, light	5
$X_{0,3}$	$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } (3) \text{ off}}$					switch the	light off	6
	$e_7 : \frac{X_{0,2} X_{2,3}}{(4)\ (2)}$					switch the	off light	7
								8
								9
								10
								11
								12
								13
								14
								15
								16
								17
								18

	apague	a	luz	por	favor	LEFT	RIGHT	NODE
$X_{1,2}$		$e_1 : \frac{\text{a}}{\text{the}}$				the	the	1
$X_{2,3}$			$e_2 : \frac{\text{luz}}{\text{light}}$			light	light	2
$X_{1,3}$		$e_3 : \frac{X_{1,2} X_{2,3}}{(1)\ (2)}$				the light	the light	3
$X_{0,2}$	$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } (1) \text{ off}}$					switch the	the off	4
$X_{2,5}$			$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } (2)}$			please ,	, light	5
$X_{0,3}$	$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } (3) \text{ off}}$					switch the	light off	6
	$e_7 : \frac{X_{0,2} X_{2,3}}{(4)\ (2)}$					switch the	off light	7
$X_{1,5}$		$e_8 : \frac{X_{1,3} \text{ por favor}}{\text{please, } (3)}$				please ,	the light	8
								9
								10
								11
								12
								13
								14
								15
								16
								17
								18

	apague a luz por favor	LEFT	RIGHT	NODE
$X_{1,2}$	$e_1 : \frac{\text{a}}{\text{the}}$	the	the	1
$X_{2,3}$	$e_2 : \frac{\text{luz}}{\text{light}}$	light	light	2
$X_{1,3}$	$e_3 : \frac{X_{1,2} X_{2,3}}{(1)\ (2)}$	the light	the light	3
$X_{0,2}$	$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } (1) \text{ off}}$	switch the	the off	4
$X_{2,5}$	$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } (2)}$	please ,	, light	5
$X_{0,3}$	$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } (3) \text{ off}}$	switch the	light off	6
	$e_7 : \frac{X_{0,2} X_{2,3}}{(4)\ (2)}$	switch the	off light	7
$X_{1,5}$	$e_8 : \frac{X_{1,3} \text{ por favor}}{\text{please, } (3)}$	please ,	the light	8
	$e_9 : \frac{X_{1,2} X_{2,5}}{(1)\ (5)}$	the please	, light	9
				10
				11
				12
				13
				14
				15
				16
				17
				18

	apague a luz por favor	LEFT	RIGHT	NODE
$X_{1,2}$	$e_1 : \frac{\text{a}}{\text{the}}$	the	the	1
$X_{2,3}$	$e_2 : \frac{\text{luz}}{\text{light}}$	light	light	2
$X_{1,3}$	$e_3 : \frac{X_{1,2} X_{2,3}}{(1)\ (2)}$	the light	the light	3
$X_{0,2}$	$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } (1) \text{ off}}$	switch the	the off	4
$X_{2,5}$	$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } (2)}$	please ,	, light	5
$X_{0,3}$	$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } (3) \text{ off}}$	switch the	light off	6
	$e_7 : \frac{X_{0,2} X_{2,3}}{(4)\ (2)}$	switch the	off light	7
$X_{1,5}$	$e_8 : \frac{X_{1,3} \text{ por favor}}{\text{please, } (3)}$	please ,	the light	8
	$e_9 : \frac{X_{1,2} X_{2,5}}{(1)\ (5)}$	the please	, light	9
$X_{0,5}$	$e_{10} : \frac{X_{0,2} X_{2,5}}{(4)\ (5)}$	switch the	, light	10
				11
				12
				13
				14
				15
				16
				17
				18

	apague a luz por favor	Left	Right	Node
$X_{1,2}$	$e_1 : \frac{\text{a}}{\text{the}}$	the	the	1
$X_{2,3}$	$e_2 : \frac{\text{luz}}{\text{light}}$	light	light	2
$X_{1,3}$	$e_3 : \frac{X_{1,2} X_{2,3}}{(1)\ (2)}$	the light	the light	3
$X_{0,2}$	$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } (1) \text{ off}}$	switch the	the off	4
$X_{2,5}$	$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } (2)}$	please ,	, light	5
$X_{0,3}$	$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } (3) \text{ off}}$	switch the	light off	6
	$e_7 : \frac{X_{0,2} X_{2,3}}{(4)\ (2)}$	switch the	off light	7
$X_{1,5}$	$e_8 : \frac{X_{1,3} \text{ por favor}}{\text{please, } (3)}$	please ,	the light	8
	$e_9 : \frac{X_{1,2} X_{2,5}}{(1)\ (5)}$	the please	, light	9
$X_{0,5}$	$e_{10} : \frac{X_{0,2} X_{2,5}}{(4)\ (5)}$	switch the	, light	10
	$e_{11} : \frac{\text{apague } X_{1,5}}{\text{switch } (8) \text{ off}}$	switch please	light off	11
				12
				13
				14
				15
				16
				17
				18

	apague a luz por favor	Left	Right	Node
$X_{1,2}$	$e_1 : \frac{\text{a}}{\text{the}}$	the	the	1
$X_{2,3}$	$e_2 : \frac{\text{luz}}{\text{light}}$	light	light	2
$X_{1,3}$	$e_3 : \frac{X_{1,2}\,X_{2,3}}{(1)\ (2)}$	the light	the light	3
$X_{0,2}$	$e_4 : \frac{\text{apague}\ X_{1,2}}{\text{switch}\ (1)\ \text{off}}$	switch the	the off	4
$X_{2,5}$	$e_5 : \frac{X_{2,3}\ \text{por favor}}{\text{please,}\ (2)}$	please ,	, light	5
$X_{0,3}$	$e_6 : \frac{\text{apague}\ X_{1,3}}{\text{switch}\ (3)\ \text{off}}$	switch the	light off	6
	$e_7 : \frac{X_{0,2}\,X_{2,3}}{(4)\ (2)}$	switch the	off light	7
$X_{1,5}$	$e_8 : \frac{X_{1,3}\ \text{por favor}}{\text{please,}\ (3)}$	please ,	the light	8
	$e_9 : \frac{X_{1,2}\,X_{2,5}}{(1)\ (5)}$	the please	, light	9
$X_{0,5}$	$e_{10} : \frac{X_{0,2}\,X_{2,5}}{(4)\ (5)}$	switch the	, light	10
	$e_{11} : \frac{\text{apague}\ X_{1,5}}{\text{switch}\ (8)\ \text{off}}$	switch please	light off	11
	$e_{12} : \frac{\text{apague}\ X_{1,5}}{\text{switch}\ (9)\ \text{off}}$	switch the	light off	12
				13
				14
				15
				16
				17
				18

	apague a luz por favor	LEFT	RIGHT	NODE
$X_{1,2}$	$e_1 : \frac{\text{a}}{\text{the}}$	the	the	1
$X_{2,3}$	$e_2 : \frac{\text{luz}}{\text{light}}$	light	light	2
$X_{1,3}$	$e_3 : \frac{X_{1,2} X_{2,3}}{(1)\ (2)}$	the light	the light	3
$X_{0,2}$	$e_4 : \frac{\text{apague}\ X_{1,2}}{\text{switch}\ (1)\ \text{off}}$	switch the	the off	4
$X_{2,5}$	$e_5 : \frac{X_{2,3}\ \text{por favor}}{\text{please,}\ (2)}$	please ,	, light	5
$X_{0,3}$	$e_6 : \frac{\text{apague}\ X_{1,3}}{\text{switch}\ (3)\ \text{off}}$	switch the	light off	6
	$e_7 : \frac{X_{0,2} X_{2,3}}{(4)\ (2)}$	switch the	off light	7
$X_{1,5}$	$e_8 : \frac{X_{1,3}\ \text{por favor}}{\text{please,}\ (3)}$	please ,	the light	8
	$e_9 : \frac{X_{1,2} X_{2,5}}{(1)\ (5)}$	the please	, light	9
$X_{0,5}$	$e_{10} : \frac{X_{0,2} X_{2,5}}{(4)\ (5)}$	switch the	, light	10
	$e_{11} : \frac{\text{apague}\ X_{1,5}}{\text{switch}\ (8)\ \text{off}}$	switch please	light off	11
	$e_{12} : \frac{\text{apague}\ X_{1,5}}{\text{switch}\ (9)\ \text{off}}$	switch the	light off	12
	$e_{13} : \frac{X_{0,3}\ \text{por favor}}{\text{please,}\ (6)}$	please ,	light off	13
				14
				15
				16
				17
				18

	apague a luz por favor	Left	Right	Node
$X_{1,2}$	$e_1 : \frac{\text{a}}{\text{the}}$	the	the	1
$X_{2,3}$	$e_2 : \frac{\text{luz}}{\text{light}}$	light	light	2
$X_{1,3}$	$e_3 : \frac{X_{1,2} X_{2,3}}{(1)\ (2)}$	the light	the light	3
$X_{0,2}$	$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } (1) \text{ off}}$	switch the	the off	4
$X_{2,5}$	$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } (2)}$	please ,	, light	5
$X_{0,3}$	$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } (3) \text{ off}}$	switch the	light off	6
	$e_7 : \frac{X_{0,2} X_{2,3}}{(4)\ (2)}$	switch the	off light	7
$X_{1,5}$	$e_8 : \frac{X_{1,3} \text{ por favor}}{\text{please, } (3)}$	please ,	the light	8
	$e_9 : \frac{X_{1,2} X_{2,5}}{(1)\ (5)}$	the please	, light	9
$X_{0,5}$	$e_{10} : \frac{X_{0,2} X_{2,5}}{(4)\ (5)}$	switch the	, light	10
	$e_{11} : \frac{\text{apague } X_{1,5}}{\text{switch } (8) \text{ off}}$	switch please	light off	11
	$e_{12} : \frac{\text{apague } X_{1,5}}{\text{switch } (9) \text{ off}}$	switch the	light off	12
	$e_{13} : \frac{X_{0,3} \text{ por favor}}{\text{please, } (6)}$	please ,	light off	13
	$e_{14} : \frac{X_{0,3} \text{ por favor}}{\text{please, } (7)}$	please ,	off light	14
				15
				16
				17
				18

	apague a luz por favor	Left	Right	Node
$X_{1,2}$	$e_1 : \frac{\text{a}}{\text{the}}$	the	the	1
$X_{2,3}$	$e_2 : \frac{\text{luz}}{\text{light}}$	light	light	2
$X_{1,3}$	$e_3 : \frac{X_{1,2}\,X_{2,3}}{(1)\ (2)}$	the light	the light	3
$X_{0,2}$	$e_4 : \frac{\text{apague}\ X_{1,2}}{\text{switch}\ (1)\ \text{off}}$	switch the	the off	4
$X_{2,5}$	$e_5 : \frac{X_{2,3}\ \text{por favor}}{\text{please,}\ (2)}$	please ,	, light	5
$X_{0,3}$	$e_6 : \frac{\text{apague}\ X_{1,3}}{\text{switch}\ (3)\ \text{off}}$	switch the	light off	6
	$e_7 : \frac{X_{0,2}\,X_{2,3}}{(4)\ (2)}$	switch the	off light	7
$X_{1,5}$	$e_8 : \frac{X_{1,3}\ \text{por favor}}{\text{please,}\ (3)}$	please ,	the light	8
	$e_9 : \frac{X_{1,2}\,X_{2,5}}{(1)\ (5)}$	the please	, light	9
$X_{0,5}$	$e_{10} : \frac{X_{0,2}\,X_{2,5}}{(4)\ (5)}$	switch the	, light	10
	$e_{11} : \frac{\text{apague}\ X_{1,5}}{\text{switch}\ (8)\ \text{off}}$	switch please	light off	11
	$e_{12} : \frac{\text{apague}\ X_{1,5}}{\text{switch}\ (9)\ \text{off}}$	switch the	light off	12
	$e_{13} : \frac{X_{0,3}\ \text{por favor}}{\text{please,}\ (6)}$	please ,	light off	13
	$e_{14} : \frac{X_{0,3}\ \text{por favor}}{\text{please,}\ (7)}$	please ,	off light	14
$S_{0,5}$	$e_{15} : \frac{\vdash X_{0,5} \dashv}{\vdash (10) \dashv}$	⊢ switch	light ⊣	15
				16
				17
				18

	apague a luz por favor	Left	Right	Node
$X_{1,2}$	$e_1 : \frac{\text{a}}{\text{the}}$	the	the	1
$X_{2,3}$	$e_2 : \frac{\text{luz}}{\text{light}}$	light	light	2
$X_{1,3}$	$e_3 : \frac{X_{1,2} X_{2,3}}{(1)\ (2)}$	the light	the light	3
$X_{0,2}$	$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } (1) \text{ off}}$	switch the	the off	4
$X_{2,5}$	$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } (2)}$	please ,	, light	5
$X_{0,3}$	$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } (3) \text{ off}}$	switch the	light off	6
	$e_7 : \frac{X_{0,2} X_{2,3}}{(4)\ (2)}$	switch the	off light	7
$X_{1,5}$	$e_8 : \frac{X_{1,3} \text{ por favor}}{\text{please, } (3)}$	please ,	the light	8
	$e_9 : \frac{X_{1,2} X_{2,5}}{(1)\ (5)}$	the please	, light	9
$X_{0,5}$	$e_{10} : \frac{X_{0,2} X_{2,5}}{(4)\ (5)}$	switch the	, light	10
	$e_{11} : \frac{\text{apague } X_{1,5}}{\text{switch } (8) \text{ off}}$	switch please	light off	11
	$e_{12} : \frac{\text{apague } X_{1,5}}{\text{switch } (9) \text{ off}}$	switch the	light off	12
	$e_{13} : \frac{X_{0,3} \text{ por favor}}{\text{please, } (6)}$	please ,	light off	13
	$e_{14} : \frac{X_{0,3} \text{ por favor}}{\text{please, } (7)}$	please ,	off light	14
$S_{0,5}$	$e_{15} : \frac{\vdash X_{0,5} \dashv}{\vdash (10) \dashv}$	⊢ switch	light ⊣	15
	$e_{16} : \frac{\vdash X_{0,5} \dashv}{\vdash (11) \dashv}$	⊢ switch	off ⊣	16
				17
				18

	apague a luz por favor	Left	Right	Node
$X_{1,2}$	$e_1 : \frac{\text{a}}{\text{the}}$	the	the	1
$X_{2,3}$	$e_2 : \frac{\text{luz}}{\text{light}}$	light	light	2
$X_{1,3}$	$e_3 : \frac{X_{1,2}\ X_{2,3}}{(1)\ \ (2)}$	the light	the light	3
$X_{0,2}$	$e_4 : \frac{\text{apague}\ X_{1,2}}{\text{switch}\ (1)\ \text{off}}$	switch the	the off	4
$X_{2,5}$	$e_5 : \frac{X_{2,3}\ \text{por favor}}{\text{please,}\ (2)}$	please ,	, light	5
$X_{0,3}$	$e_6 : \frac{\text{apague}\ X_{1,3}}{\text{switch}\ (3)\ \text{off}}$	switch the	light off	6
	$e_7 : \frac{X_{0,2}\ X_{2,3}}{(4)\ \ (2)}$	switch the	off light	7
$X_{1,5}$	$e_8 : \frac{X_{1,3}\ \text{por favor}}{\text{please,}\ (3)}$	please ,	the light	8
	$e_9 : \frac{X_{1,2}\ X_{2,5}}{(1)\ \ (5)}$	the please	, light	9
$X_{0,5}$	$e_{10} : \frac{X_{0,2}\ X_{2,5}}{(4)\ \ (5)}$	switch the	, light	10
	$e_{11} : \frac{\text{apague}\ X_{1,5}}{\text{switch}\ (8)\ \text{off}}$	switch please	light off	11
	$e_{12} : \frac{\text{apague}\ X_{1,5}}{\text{switch}\ (9)\ \text{off}}$	switch the	light off	12
	$e_{13} : \frac{X_{0,3}\ \text{por favor}}{\text{please,}\ (6)}$	please ,	light off	13
	$e_{14} : \frac{X_{0,3}\ \text{por favor}}{\text{please,}\ (7)}$	please ,	off light	14
$S_{0,5}$	$e_{15} : \frac{\vdash X_{0,5} \dashv}{\vdash (10) \dashv}$	$\vdash$ switch	light $\dashv$	15
	$e_{16} : \frac{\vdash X_{0,5} \dashv}{\vdash (11) \dashv}$ or $e_{17} : \frac{\vdash X_{0,5} \dashv}{\vdash (12) \dashv}$	$\vdash$ switch	off $\dashv$	16
				17
				18

	apague a luz por favor	Left	Right	Node
$X_{1,2}$	$e_1 : \frac{\text{a}}{\text{the}}$	the	the	1
$X_{2,3}$	$e_2 : \frac{\text{luz}}{\text{light}}$	light	light	2
$X_{1,3}$	$e_3 : \frac{X_{1,2}X_{2,3}}{(1)\ (2)}$	the light	the light	3
$X_{0,2}$	$e_4 : \frac{\text{apague } X_{1,2}}{\text{switch } (1) \text{ off}}$	switch the	the off	4
$X_{2,5}$	$e_5 : \frac{X_{2,3} \text{ por favor}}{\text{please, } (2)}$	please ,	, light	5
$X_{0,3}$	$e_6 : \frac{\text{apague } X_{1,3}}{\text{switch } (3) \text{ off}}$	switch the	light off	6
	$e_7 : \frac{X_{0,2}X_{2,3}}{(4)\ (2)}$	switch the	off light	7
$X_{1,5}$	$e_8 : \frac{X_{1,3} \text{ por favor}}{\text{please, } (3)}$	please ,	the light	8
	$e_9 : \frac{X_{1,2}X_{2,5}}{(1)\ (5)}$	the please	, light	9
$X_{0,5}$	$e_{10} : \frac{X_{0,2}X_{2,5}}{(4)\ (5)}$	switch the	, light	10
	$e_{11} : \frac{\text{apague } X_{1,5}}{\text{switch } (8) \text{ off}}$	switch please	light off	11
	$e_{12} : \frac{\text{apague } X_{1,5}}{\text{switch } (9) \text{ off}}$	switch the	light off	12
	$e_{13} : \frac{X_{0,3} \text{ por favor}}{\text{please, } (6)}$	please ,	light off	13
	$e_{14} : \frac{X_{0,3} \text{ por favor}}{\text{please, } (7)}$	please ,	off light	14
$S_{0,5}$	$e_{15} : \frac{\vdash X_{0,5} \dashv}{\vdash (10) \dashv}$	$\vdash$ switch	light $\dashv$	15
	$e_{16} : \frac{\vdash X_{0,5} \dashv}{\vdash (11) \dashv}$ or $e_{17} : \frac{\vdash X_{0,5} \dashv}{\vdash (12) \dashv}$	$\vdash$ switch	off $\dashv$	16
	$e_{18} : \frac{\vdash X_{0,5} \dashv}{\vdash (13) \dashv}$	$\vdash$ please	off $\dashv$	17
				18

	apague a luz por favor	Left	Right	Node
$X_{1,2}$	$e_1 : \frac{\text{a}}{\text{the}}$	the	the	1
$X_{2,3}$	$e_2 : \frac{\text{luz}}{\text{light}}$	light	light	2
$X_{1,3}$	$e_3 : \frac{X_{1,2}X_{2,3}}{(1)\ (2)}$	the light	the light	3
$X_{0,2}$	$e_4 : \frac{\text{apague}\ X_{1,2}}{\text{switch}\ (1)\ \text{off}}$	switch the	the off	4
$X_{2,5}$	$e_5 : \frac{X_{2,3}\ \text{por favor}}{\text{please,}\ (2)}$	please ,	, light	5
$X_{0,3}$	$e_6 : \frac{\text{apague}\ X_{1,3}}{\text{switch}\ (3)\ \text{off}}$	switch the	light off	6
	$e_7 : \frac{X_{0,2}X_{2,3}}{(4)\ (2)}$	switch the	off light	7
$X_{1,5}$	$e_8 : \frac{X_{1,3}\ \text{por favor}}{\text{please,}\ (3)}$	please ,	the light	8
	$e_9 : \frac{X_{1,2}X_{2,5}}{(1)\ (5)}$	the please	, light	9
$X_{0,5}$	$e_{10} : \frac{X_{0,2}X_{2,5}}{(4)\ (5)}$	switch the	, light	10
	$e_{11} : \frac{\text{apague}\ X_{1,5}}{\text{switch}\ (8)\ \text{off}}$	switch please	light off	11
	$e_{12} : \frac{\text{apague}\ X_{1,5}}{\text{switch}\ (9)\ \text{off}}$	switch the	light off	12
	$e_{13} : \frac{X_{0,3}\ \text{por favor}}{\text{please,}\ (6)}$	please ,	light off	13
	$e_{14} : \frac{X_{0,3}\ \text{por favor}}{\text{please,}\ (7)}$	please ,	off light	14
$S_{0,5}$	$e_{15} : \frac{\vdash X_{0,5} \dashv}{\vdash (10) \dashv}$	$\vdash$ switch	light $\dashv$	15
	$e_{16} : \frac{\vdash X_{0,5} \dashv}{\vdash (11) \dashv}$ or $e_{17} : \frac{\vdash X_{0,5} \dashv}{\vdash (12) \dashv}$	$\vdash$ switch	off $\dashv$	16
	$e_{18} : \frac{\vdash X_{0,5} \dashv}{\vdash (13) \dashv}$	$\vdash$ please	off $\dashv$	17
	$e_{19} : \frac{\vdash X_{0,5} \dashv}{\vdash (14) \dashv}$	$\vdash$ please	light $\dashv$	18

The problem with the solution

The problem with the solution

Computational complexity!

The problem with the solution

Computational complexity!

1. it seems like the underlying grammar is growing
2. there are way too many n-grams leading to way too many nonterminals
3. the graphical representation (forest) is growing

What is really going on?

We are transferring memory from an automaton to the forest

What is really going on?

We are transferring memory from an automaton to the forest

- n-gram LMs can be thought of as an automaton where each state q uniquely represents a k-gram prefix α_q $(k < n)$

What is really going on?

We are transferring memory from an automaton to the forest

- n-gram LMs can be thought of as an automaton where each state q uniquely represents a k-gram prefix α_q $(k < n)$
- a transition from a state q labelled with word w is weighted by the probability $p(w|\alpha_q)$

What is really going on?

We are transferring memory from an automaton to the forest

- n-gram LMs can be thought of as an automaton where each state q uniquely represents a k-gram prefix α_q $(k < n)$
- a transition from a state q labelled with word w is weighted by the probability $p(w|\alpha_q)$

A nonterminal in the forest yields a set of strings

What is really going on?

We are transferring memory from an automaton to the forest

- n-gram LMs can be thought of as an automaton where each state q uniquely represents a k-gram prefix α_q $(k < n)$
- a transition from a state q labelled with word w is weighted by the probability $p(w|\alpha_q)$

A nonterminal in the forest yields a set of strings

- strings project onto paths in the LM automaton

What is really going on?

We are transferring memory from an automaton to the forest

- n-gram LMs can be thought of as an automaton where each state q uniquely represents a k-gram prefix α_q $(k < n)$
- a transition from a state q labelled with word w is weighted by the probability $p(w|\alpha_q)$

A nonterminal in the forest yields a set of strings

- strings project onto paths in the LM automaton
- paths are weighted

What is really going on?

We are transferring memory from an automaton to the forest

- n-gram LMs can be thought of as an automaton where each state q uniquely represents a k-gram prefix α_q $(k < n)$
- a transition from a state q labelled with word w is weighted by the probability $p(w|\alpha_q)$

A nonterminal in the forest yields a set of strings

- strings project onto paths in the LM automaton
- paths are weighted

Nonterminals must be aware of (parts of) the strings they yield

What is really going on?

We are transferring memory from an automaton to the forest

- n-gram LMs can be thought of as an automaton where each state q uniquely represents a k-gram prefix α_q $(k < n)$
- a transition from a state q labelled with word w is weighted by the probability $p(w|\alpha_q)$

A nonterminal in the forest yields a set of strings

- strings project onto paths in the LM automaton
- paths are weighted

Nonterminals must be aware of (parts of) the strings they yield

- they must be annotated with states of the automaton

How hard is it?

Weighted intersection between a wCFG and a wFSA

How hard is it?

Weighted intersection between a wCFG and a wFSA

Generalisation of parsing for

How hard is it?

Weighted intersection between a wCFG and a wFSA

Generalisation of parsing for

- arbitrary automata

How hard is it?

Weighted intersection between a wCFG and a wFSA

Generalisation of parsing for

- arbitrary automata
- weighted sets

How hard is it?

Weighted intersection between a wCFG and a wFSA

Generalisation of parsing for

- arbitrary automata
- weighted sets

Complexity

How hard is it?

Weighted intersection between a wCFG and a wFSA

Generalisation of parsing for

- arbitrary automata
- weighted sets

Complexity

- input: $X_0 \rightarrow X_1 X_2 \dots X_a$ where $X_i \in N$

How hard is it?

Weighted intersection between a wCFG and a wFSA

Generalisation of parsing for

- arbitrary automata
- weighted sets

Complexity

- input: $X_0 \rightarrow X_1 X_2 \ldots X_a$ where $X_i \in N$
- output: $X_0^{(q_1,q_a)} \rightarrow X_1^{(q_1,q_2)} X_2^{(q_2,q_3)} \ldots X_a^{(q_{a-1},q_a)}$ where $q_j \in Q$

How hard is it?

Weighted intersection between a wCFG and a wFSA

Generalisation of parsing for

- arbitrary automata
- weighted sets

Complexity

- input: $X_0 \rightarrow X_1 X_2 \dots X_a$ where $X_i \in N$
- output: $X_0^{(q_1,q_a)} \rightarrow X_1^{(q_1,q_2)} X_2^{(q_2,q_3)} \dots X_a^{(q_{a-1},q_a)}$
  where $q_j \in Q$
- complexity: $O(|N||Q|^{a+1})$

Solution

The usual suspect

- pruning

Solution

The usual suspect

- pruning

Alternatives

- local search (greedy methods)
- relaxation techniques
- sampling

Pruning: beam search

Approximate intersection by budgeting the combination of "comparable" nodes

Pruning: beam search

Approximate intersection by budgeting the combination of "comparable" nodes

- nodes that share structure

Pruning: beam search

Approximate intersection by budgeting the combination of "comparable" nodes

- nodes that share structure
    - phrase-based: coverage vector
    - hierarchical: input spans

Pruning: beam search

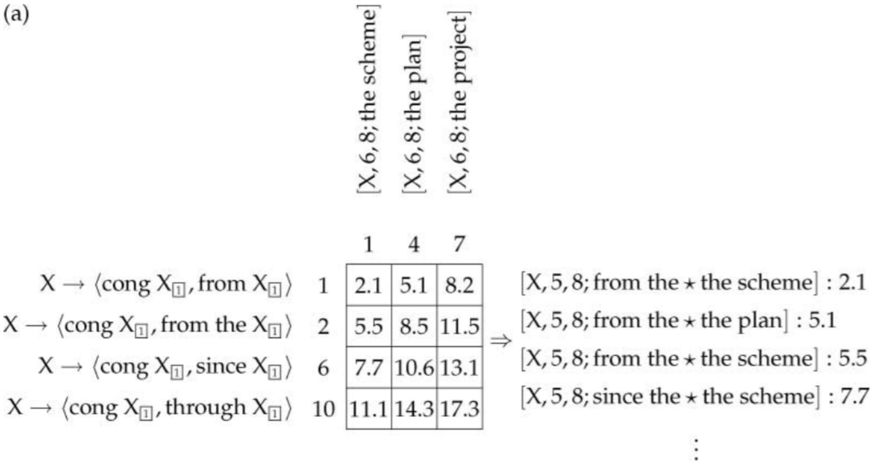
Approximate intersection by budgeting the combination of "comparable" nodes

- nodes that share structure
  - phrase-based: coverage vector
  - hierarchical: input spans
- heuristic view of interaction with LM
  - phrase-based: approximate future cost
  - local approximation based on limited context

Naive beam search

1. enumerate combinations
2. sort and prune all but the k best

Naive beam search

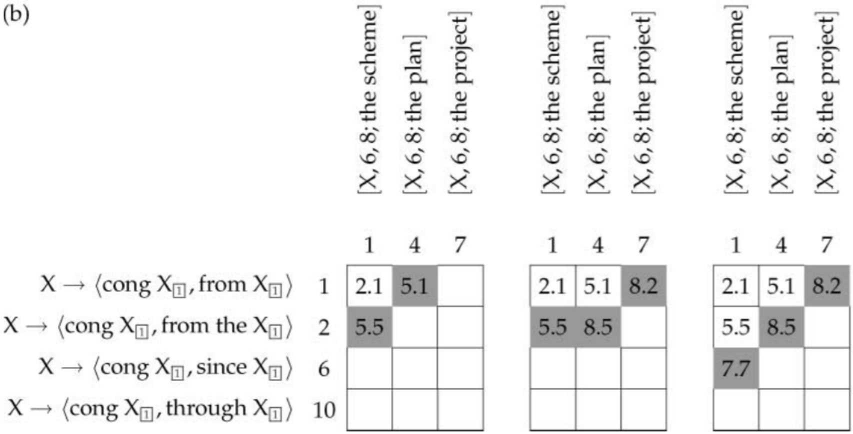


Cube pruning

An agenda for pruning [Chiang, 2007]

- tries to enumerate combinations in best-first order
- stops after k items have been enumerated
- inspiration: product of sorted lists
- heuristic: assumes the LM a monotone function over edges

Cube pruning



Problem with pruning

1. unbounded approximation
2. approximating the Viterbi solution
3. incompatible with models which have a probabilistic interpretation
4. cannot handle arbitrarily nonlocal dependencies

Beyond beam search

Local search [Hardmeier et al., 2012]

- computationally cheap
- unbounded approximation
- approximate Viterbi
- can handle arbitrarily nonlocal dependencies
- too local view of the distribution (bad for tuning)

Beyond beam search

Relaxation methods
[Chang and Collins, 2011, Rush and Collins, 2011]

- computationally expensive
- bounded approximation
- (approximate) Viterbi
- may handle arbitrarily nonlocal dependencies

Beyond beam search

Sampling [Arun et al., 2009, Aziz et al., 2013, Aziz, 2014]

- (bounded) approximation
- (approximate) Viterbi, expectations
- handle arbitrarily nonlocal dependencies
- in principle ideal for tuning (global view of distribution)
- potentially computationally expensive

Questions?

References I

Abhishek Arun, Chris Dyer, Barry Haddow, Phil Blunsom, Adam Lopez, and Philipp Koehn. Monte Carlo inference and maximization for phrase-based translation. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 102–110, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-29-9. URL http://dl.acm.org/citation.cfm?id=1596374.1596394.

Wilker Aziz, Marc Dymetman, and Sriram Venkatapathy. Investigations in exact inference for hierarchical translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 472–483, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W13-2260.

References II

Wilker Ferreira Aziz. *Exact Sampling and Optimisation in Statistical Machine Translation*. PhD thesis, University of Wolverhampton, 2014.

Yin-Wen Chang and Michael Collins. Exact decoding of phrase-based translation models through Lagrangian relaxation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 26–37, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-937284-11-4.

David Chiang. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228, June 2007. ISSN 0891-2017. doi: 10.1162/coli.2007.33.2.201. URL http://dx.doi.org/10.1162/coli.2007.33.2.201.

References III

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. Document-wide decoding for phrase-based statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1179–1190, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D12-1108.

Alexander M. Rush and Michael Collins. Exact decoding of syntactic translation models through Lagrangian relaxation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 72–82, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-87-9.