

Machine Translation

Luciano Barbosa

(slides baseados no curso de NLP de Stanford)

Definição

- Tarefa de traduzir um sentença x de uma língua (source language) para uma sentença y em uma outra língua (target language)

x: *L'homme est né libre, et partout il est dans les fers*



y: *Man is born free, but everywhere he is in chains*

Início da Área

- Anos de 1950
- Motivada pela guerra fria
 - Russo -> Inglês
 - Baseado em regras, usando dicionário bilíngüe

Statistical Machine Translation

- Ideia básica: aprender um modelo probabilístico a partir dos dados
- Ex: melhor sentença y em inglês, dada uma sentença x em francês

$$\operatorname{argmax}_y P(y|x)$$

- Regra de Bayes quebra em 2 componentes:

$$= \operatorname{argmax}_y \underbrace{P(x|y)}_{\text{Translation Model}} \underbrace{P(y)}_{\text{Language Model}}$$

Translation Model

Models how words and phrases should be translated (*fidelity*).
Learnt from parallel data.

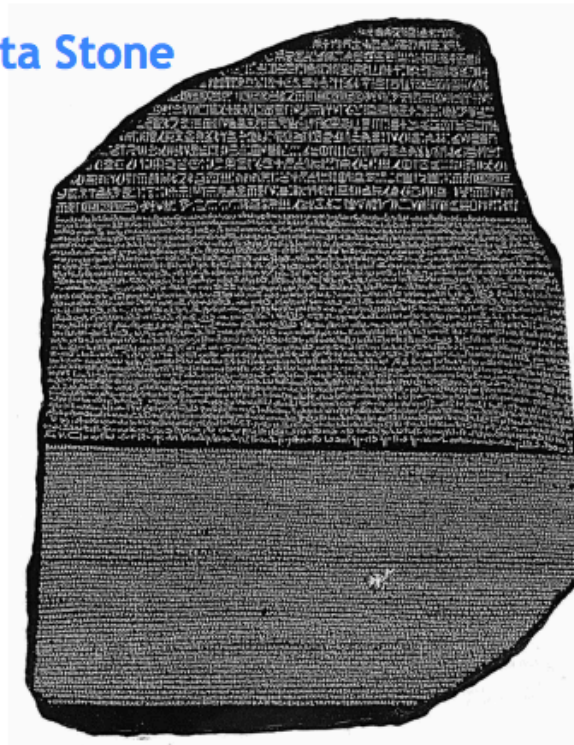
Language Model

Models how to write good English (*fluency*).
Learnt from monolingual data.

Statistical Machine Translation

- Como aprender o modelo de tradução $P(x|y)$?
- Grande quantidade de corpus paralelo

The Rosetta Stone



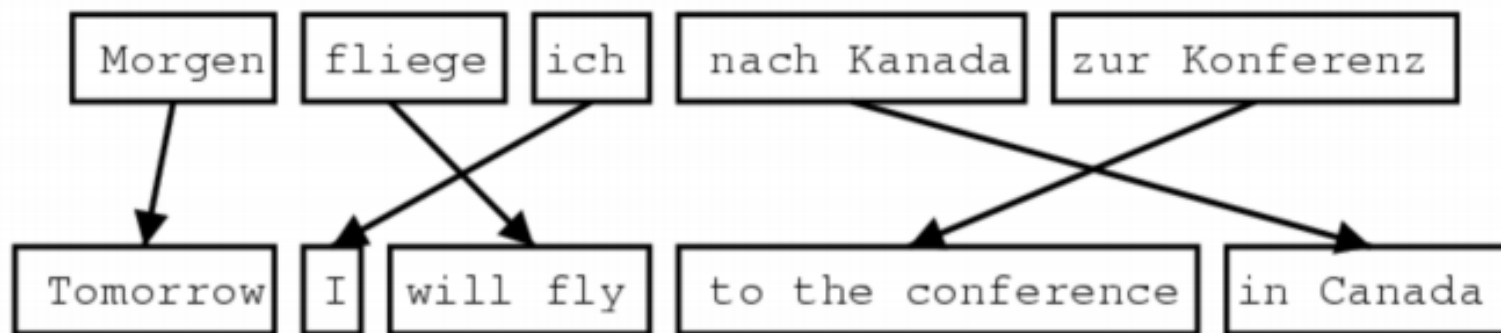
Ancient Egyptian

Demotic

Ancient Greek

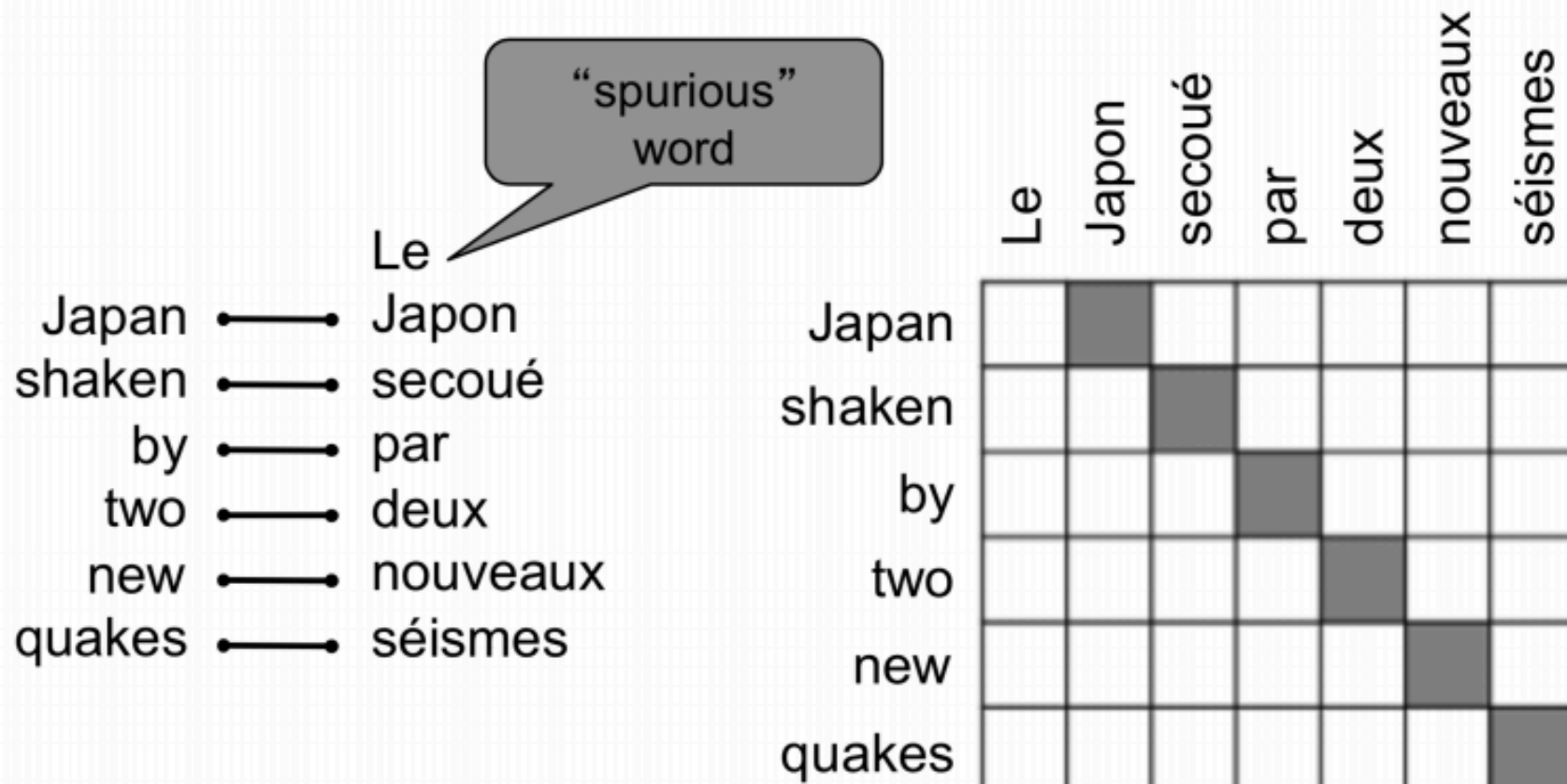
Alinhamento de Corpus Paralelo

- Definir alinhamentos entre sentenças paralelas



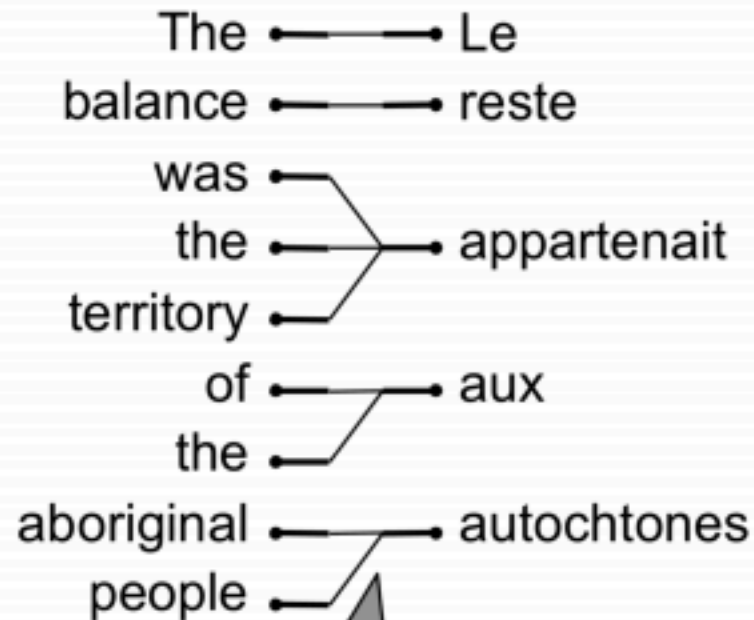
- Desafios:
 - Diferenças na construção de sentenças em diferentes línguas
 - Algumas palavras não possuem tradução

Exemplos



Examples from: “The Mathematics of Statistical Machine Translation: Parameter Estimation”, Brown et al, 1993. <http://www.aclweb.org/anthology/J93-2003>

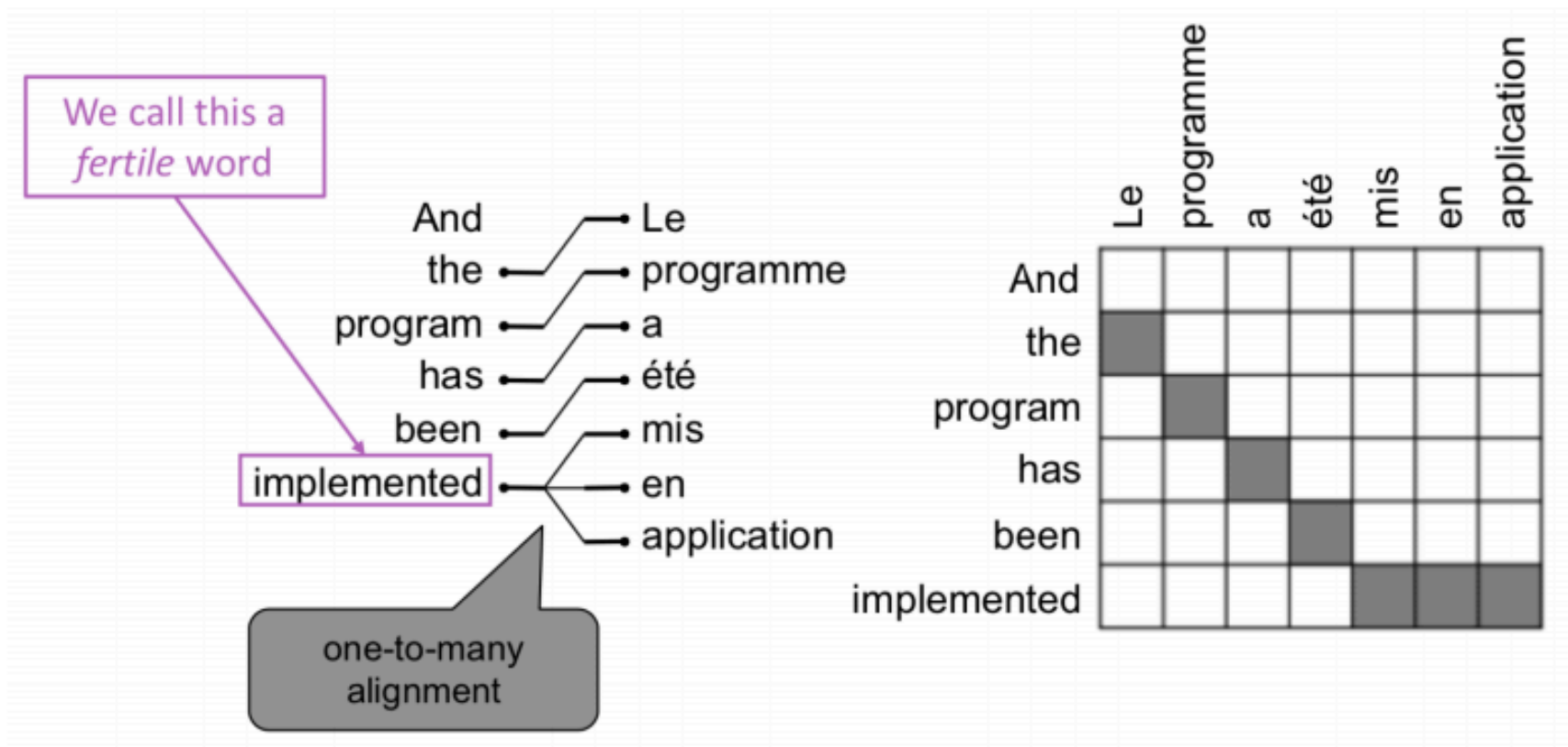
Exemplos



many-to-one
alignments

	Le	reste	appartenait	aux	autochtones
The					
balance					
was					
the					
territory					
of					
the					
aboriginal					
people					

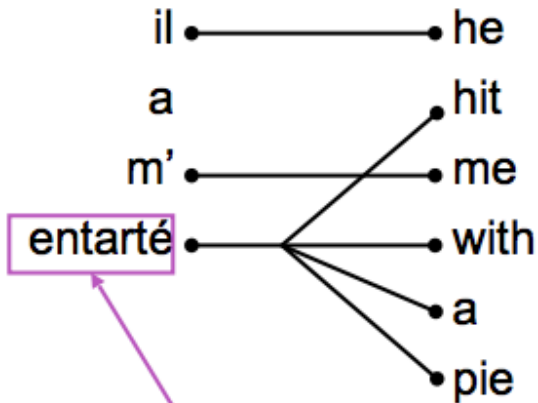
Exemplos



Exemplos

he hit me with a pie

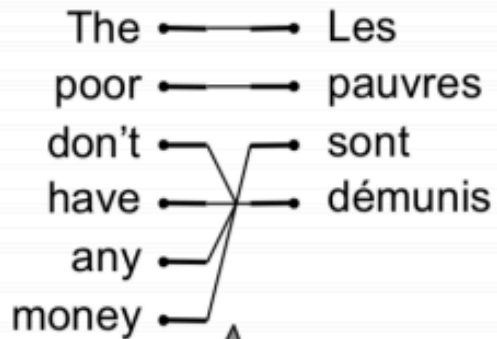
il					
a					
m'					
entarté					



This word has no single-
 word equivalent in
 English



Exemplos



many-to-many
alignment

	Les	pauvres	sont	démunis
The				
poor				
don't				
have				
any				
money				

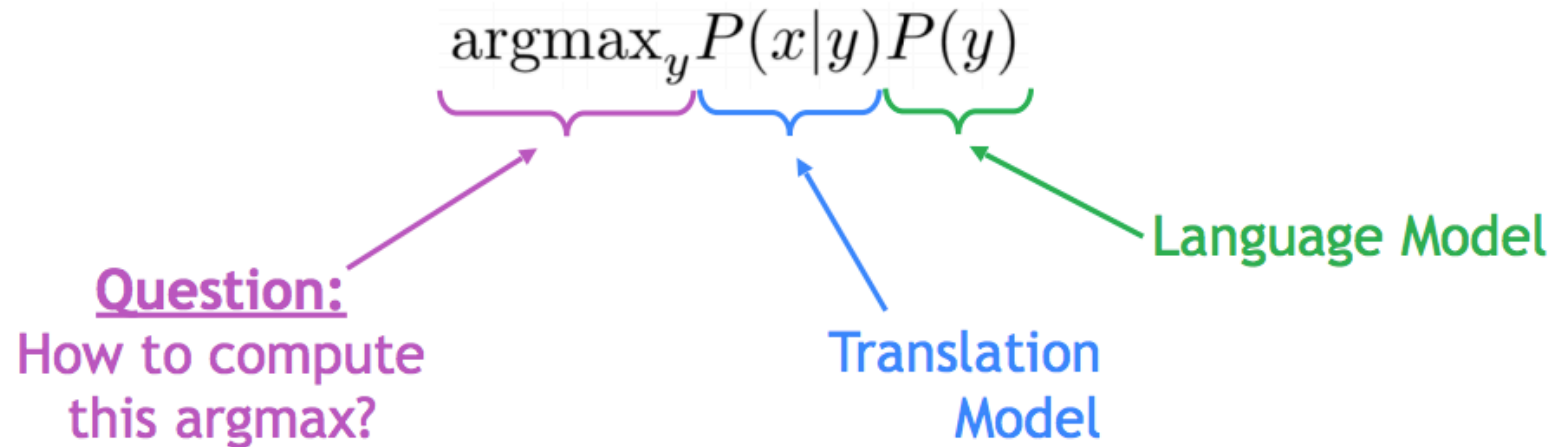
phrase
alignment

Examples from: "The Mathematics of Statistical Machine Translation: Parameter Estimation", Brown et al, 1993. <http://www.aclweb.org/anthology/J93-2003>

Aprendendo o Alinhamento

- Combinação de fatores
 - Probabilidade do alinhamento de palavras específicas (depende da posição na sentença)
 - Probabilidade de uma palavra ter fertilidade
 - etc
- Utilização de algoritmos como Expectation-Maximization

Decoding



○ Que hambre tengo yo

argmax_e	What hunger have I	$p(s e)p(e) = 0.000014 \times 0.000001$	$\left. \vphantom{\begin{matrix} \text{What hunger have I} \\ \text{Hungry I am so} \\ \text{I am so hungry} \\ \text{Have I that hunger} \\ \dots \end{matrix}} \right\} = \text{I am so hungry}$
	Hungry I am so	$p(s e)p(e) = 0.000001 \times 0.0000014$	
	I am so hungry	$p(s e)p(e) = 0.0000015 \times 0.0001$	
	Have I that hunger	$p(s e)p(e) = 0.000020 \times 0.00000098$	
	...		

Statistical Machine Translation

- Melhores sistemas bem complexos
- Feature engineering para capturar detalhes das linguagens
- Muito esforço humano para manter

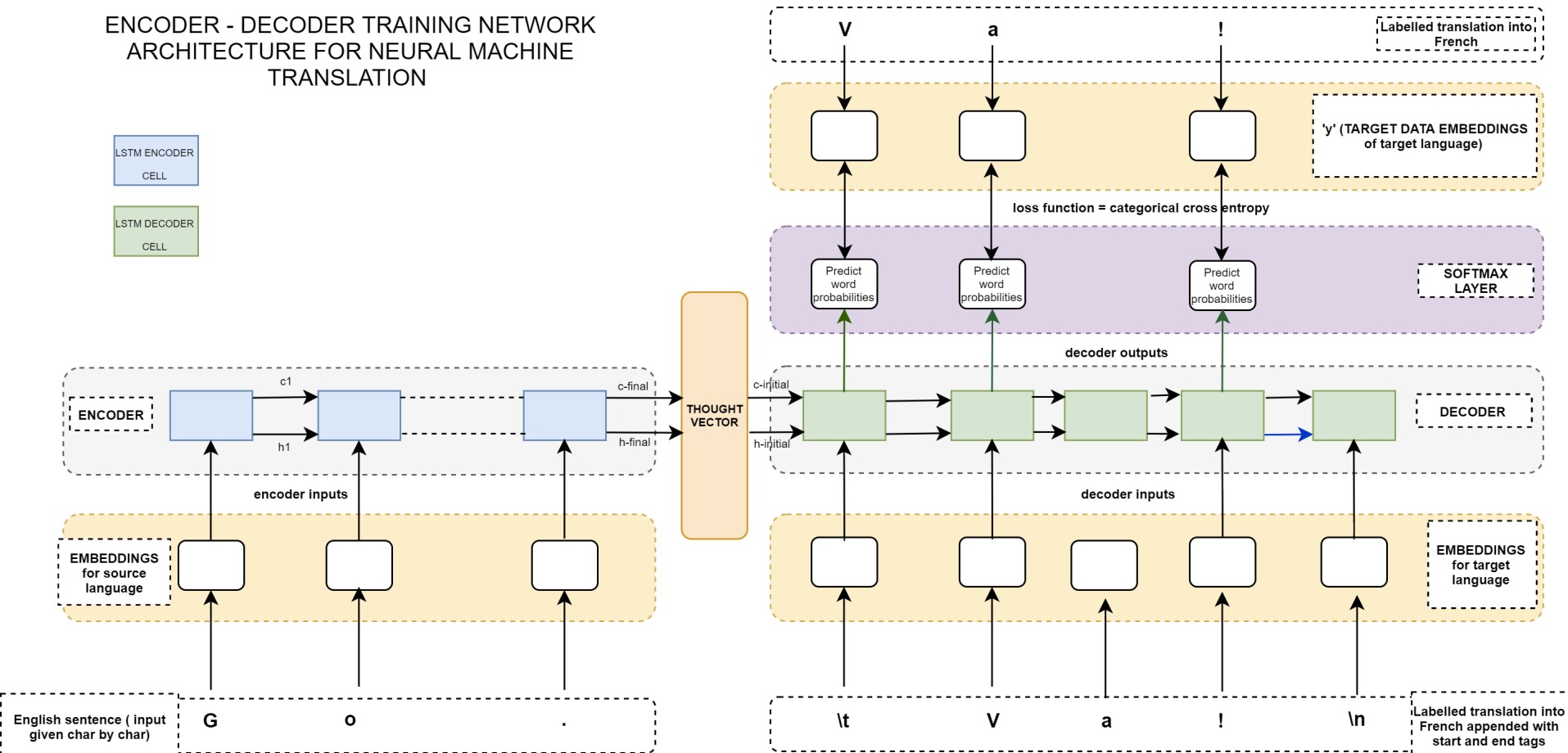
Neural Machine Translation

- Usa uma única rede neural
- Arquitetura: sequence-to-sequence
 - Utilizada também para: diálogo, parsing, geração de texto
- Duas etapas:
 - Treinamento
 - Tradução

Treinamento

- Precisa de corpus paralelo

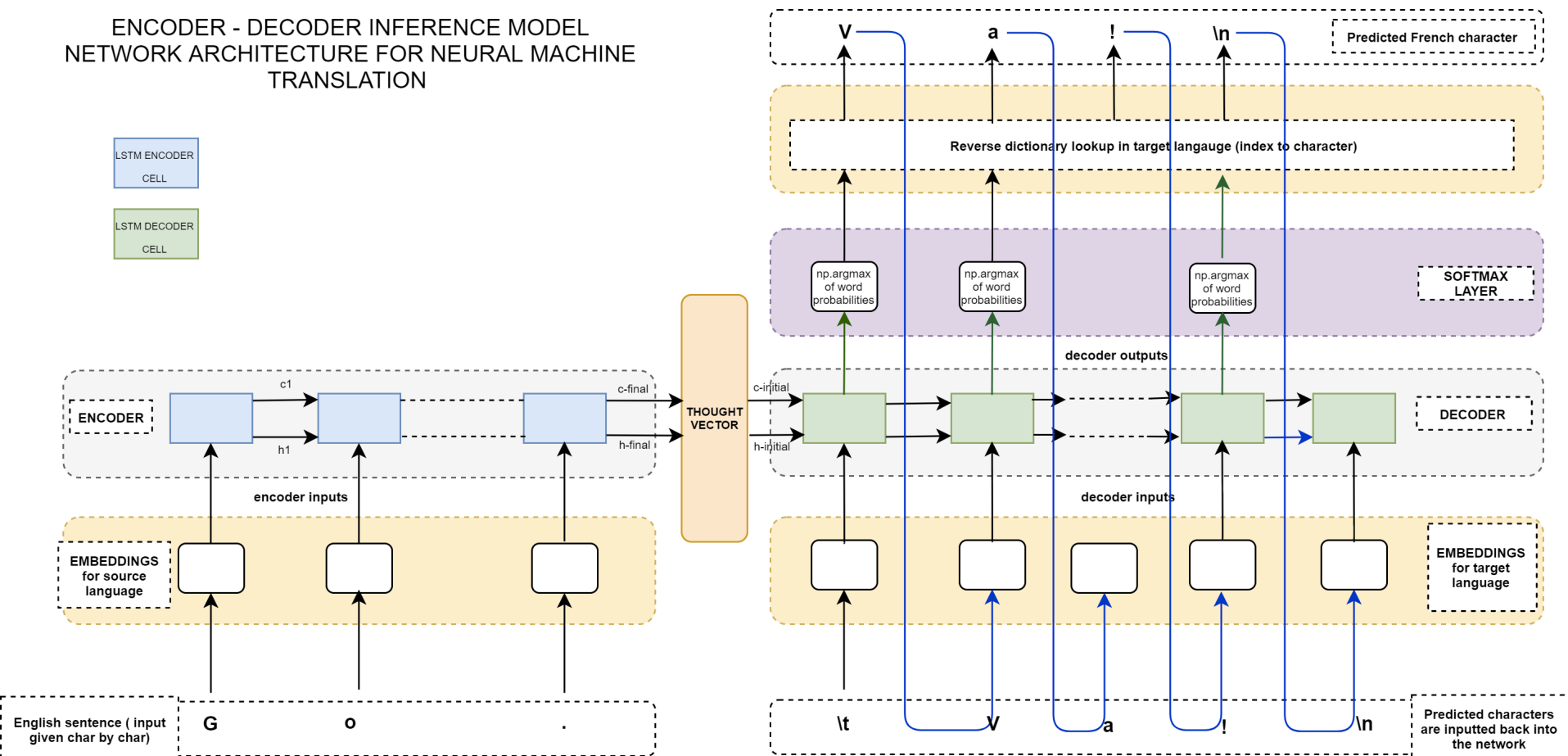
ENCODER - DECODER TRAINING NETWORK ARCHITECTURE FOR NEURAL MACHINE TRANSLATION



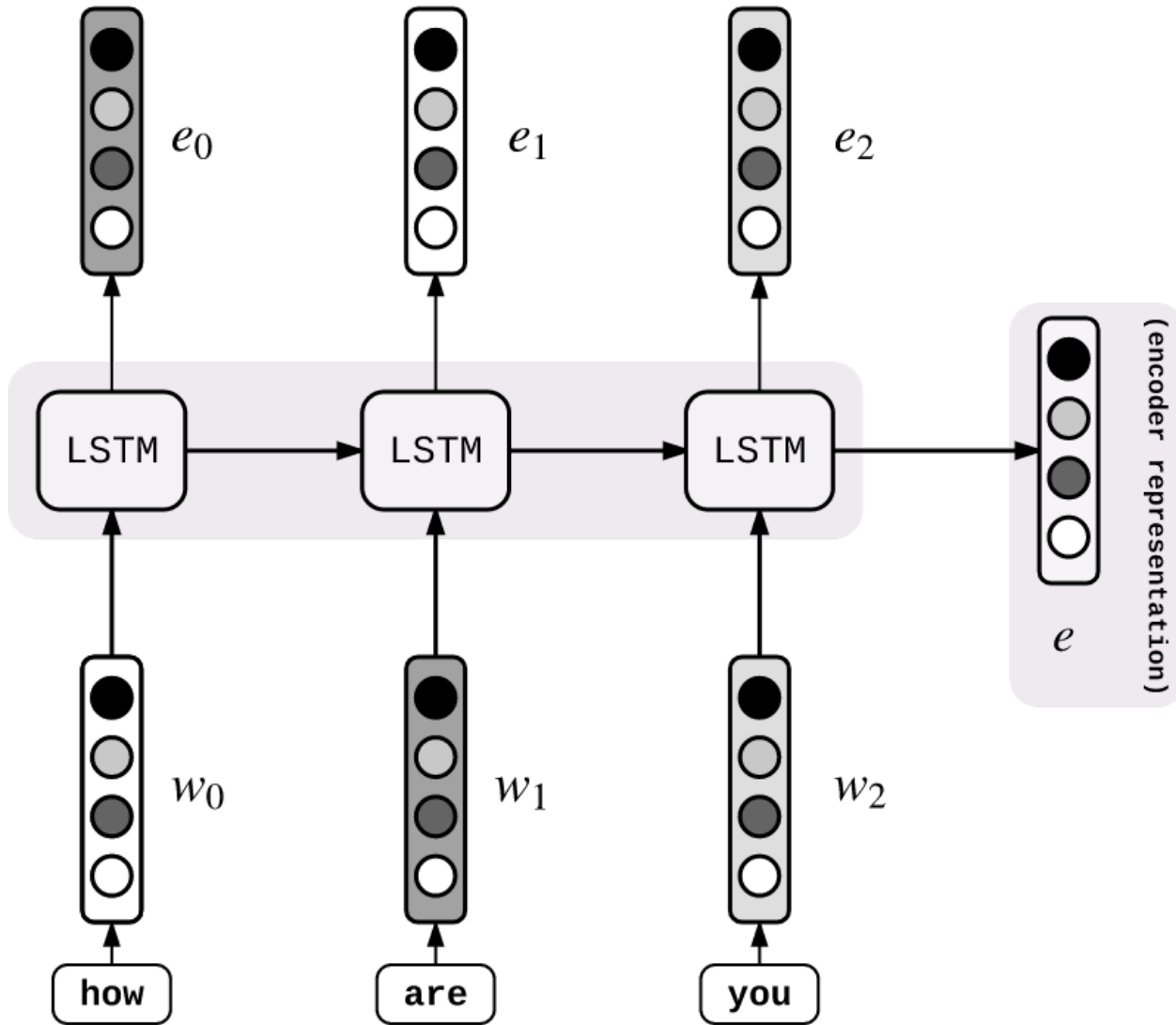
Tradução

- Greedy: Seleciona a palavra com máxima probabilidade

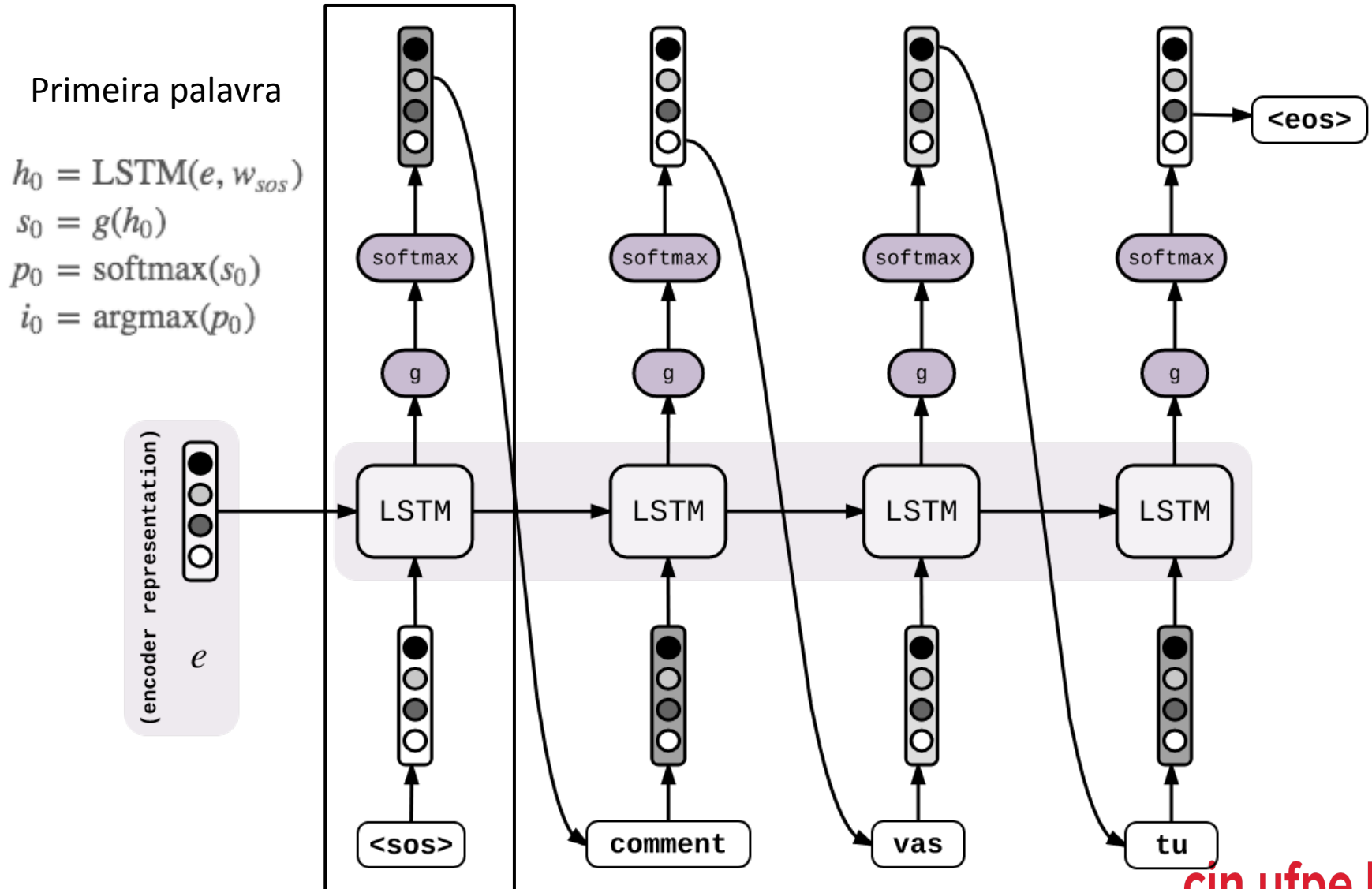
ENCODER - DECODER INFERENCE MODEL
NETWORK ARCHITECTURE FOR NEURAL MACHINE
TRANSLATION



Encoder



Decoder (Greedy)



Decoder (Greedy)

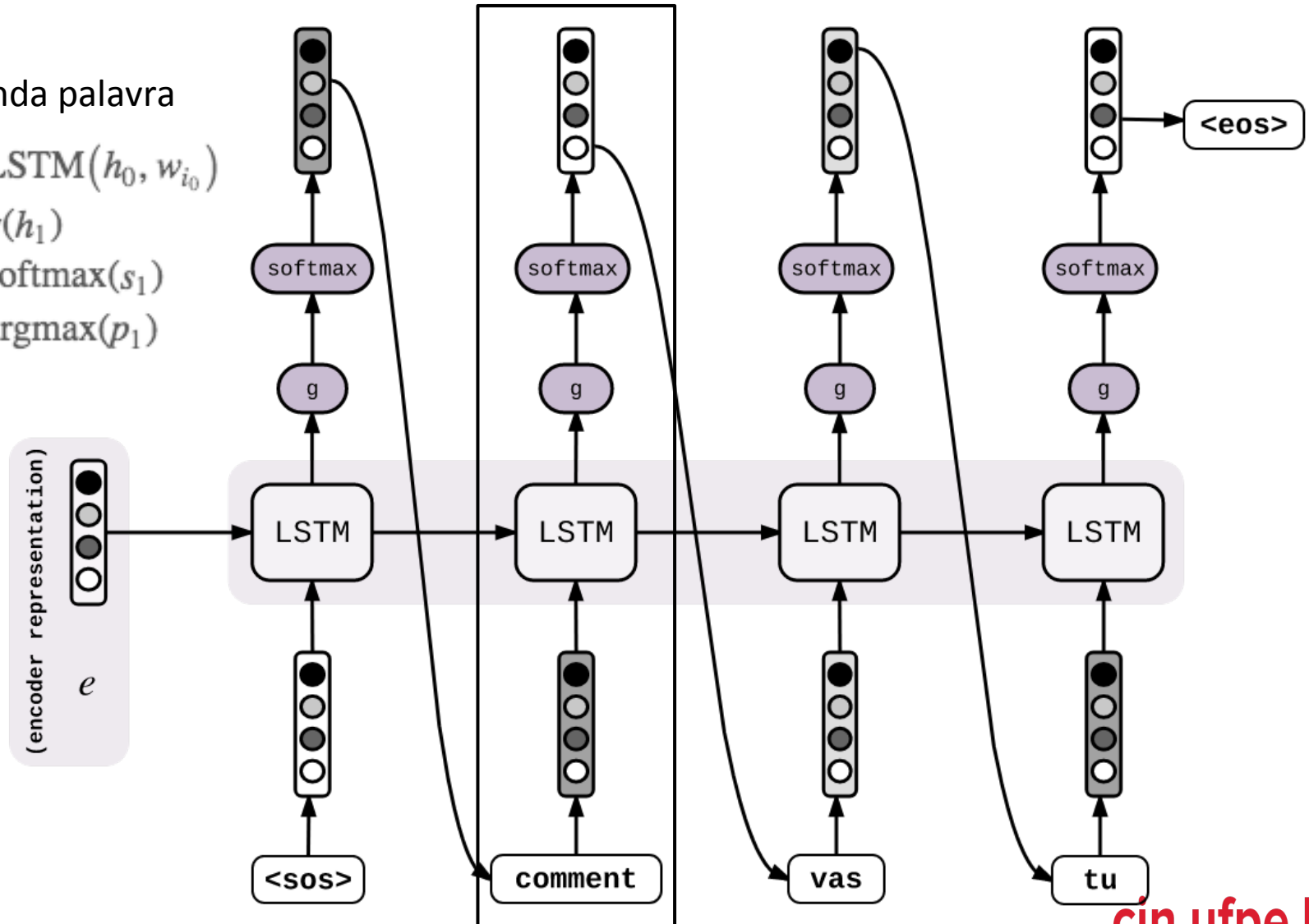
Segunda palavra

$$h_1 = \text{LSTM}(h_0, w_{i_0})$$

$$s_1 = g(h_1)$$

$$p_1 = \text{softmax}(s_1)$$

$$i_1 = \text{argmax}(p_1)$$

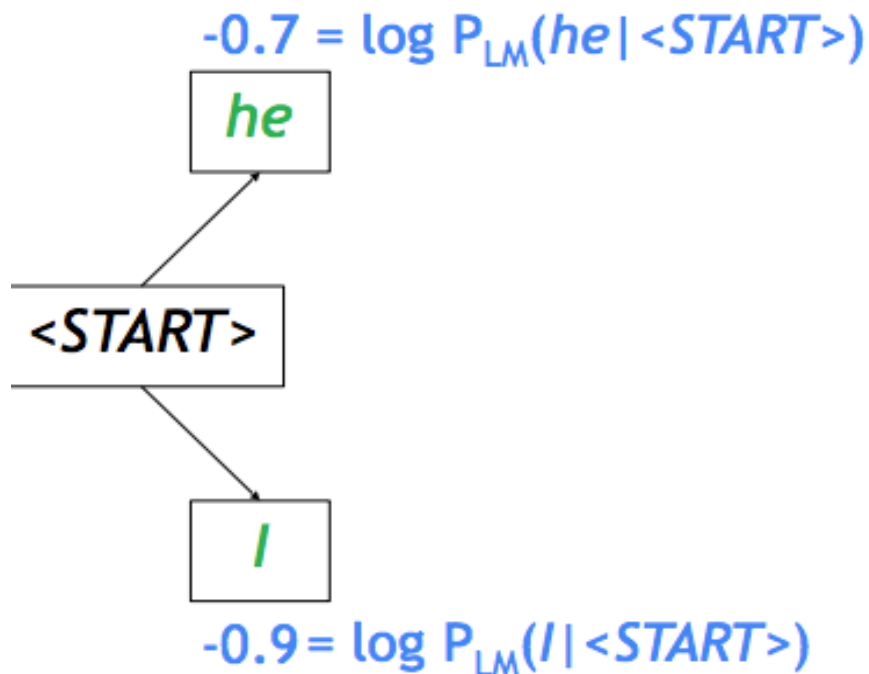


Beam Search

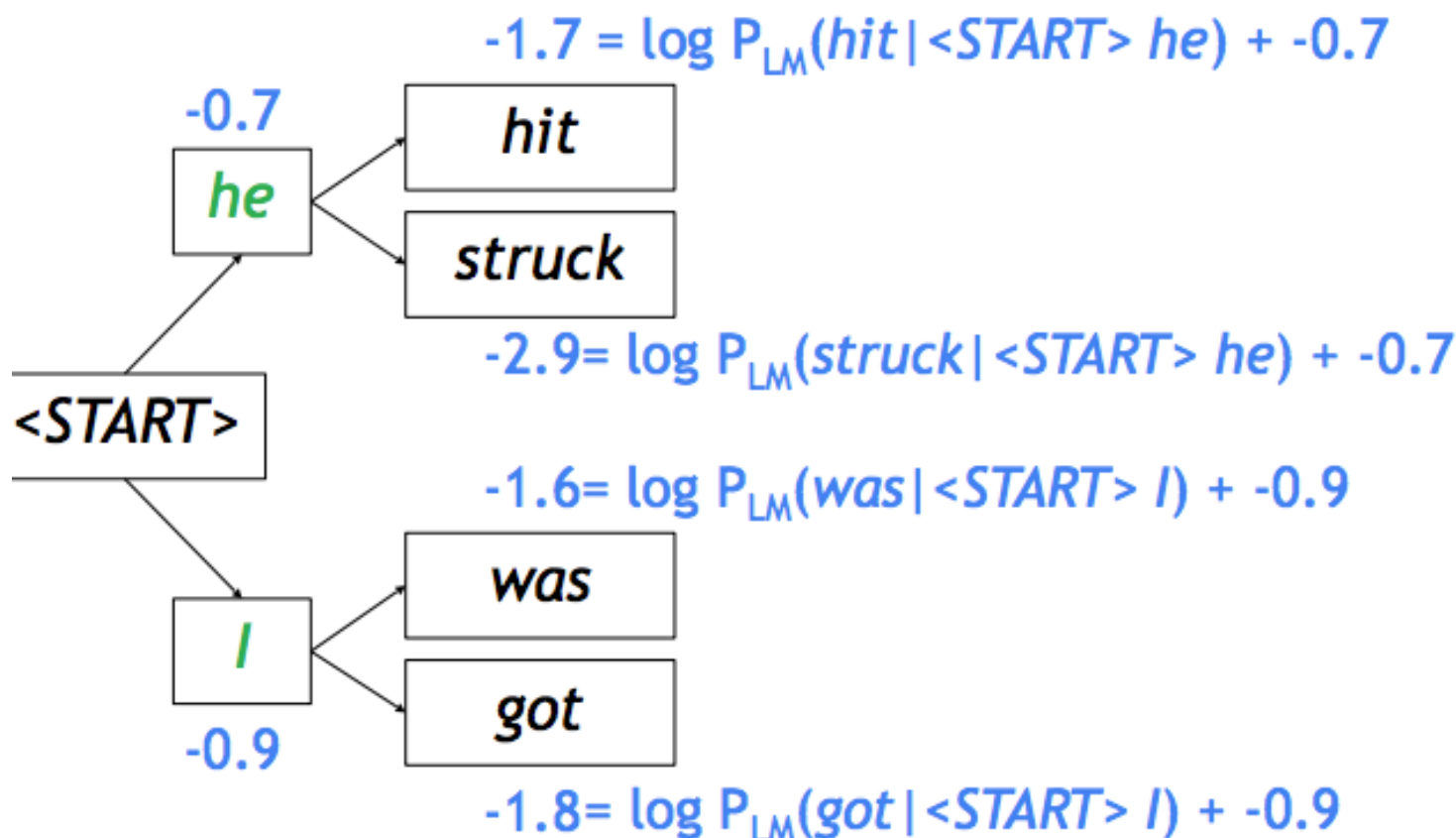
- Limitação do greedy decoding: não permite desfazer decisões
- Em cada passo, manter as k traduções parciais mais prováveis ou hipóteses
- Não é garantido encontrar a melhor solução

Beam Search: Example

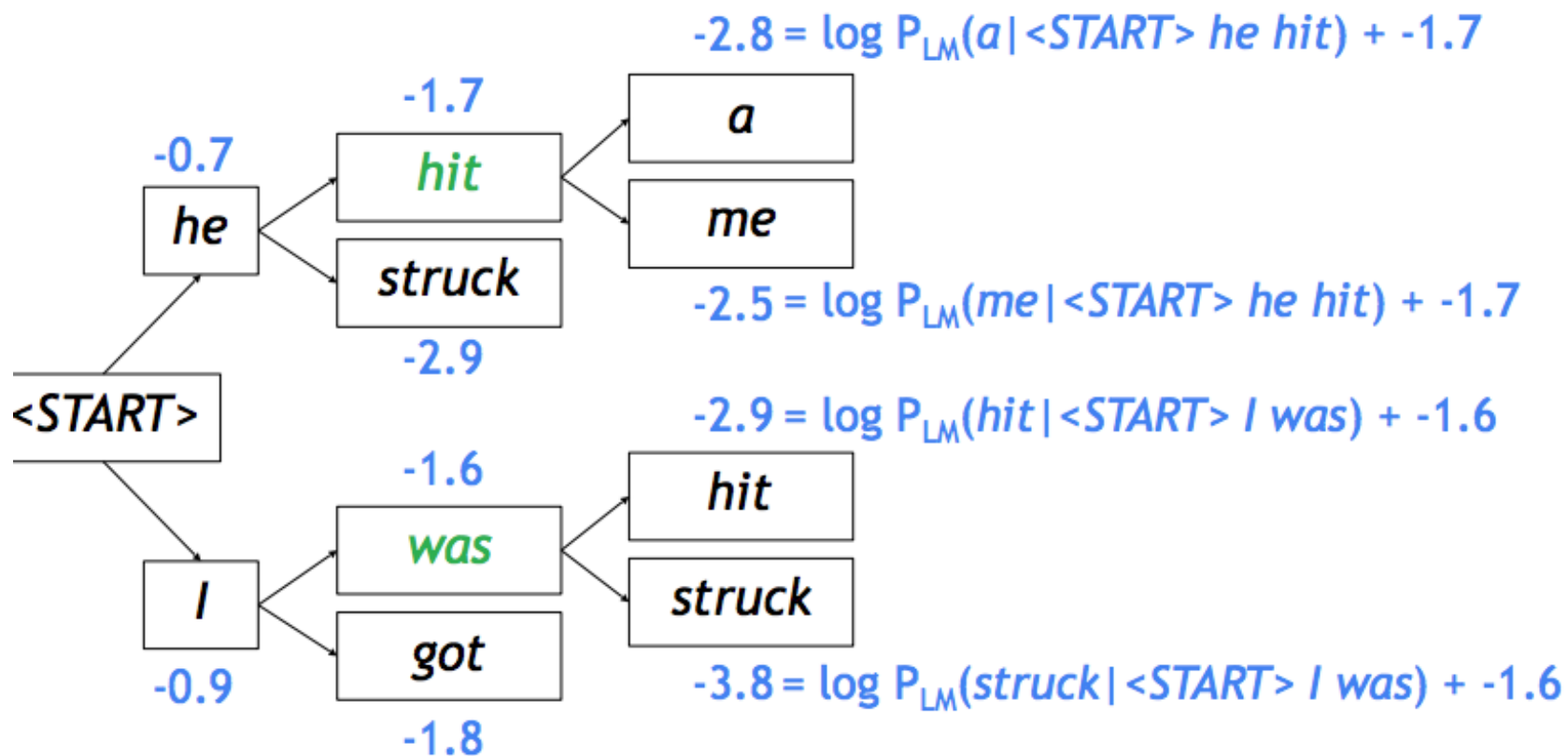
- Source: il a m' entarté
- Target: he hit me with a pie
- $K=2$



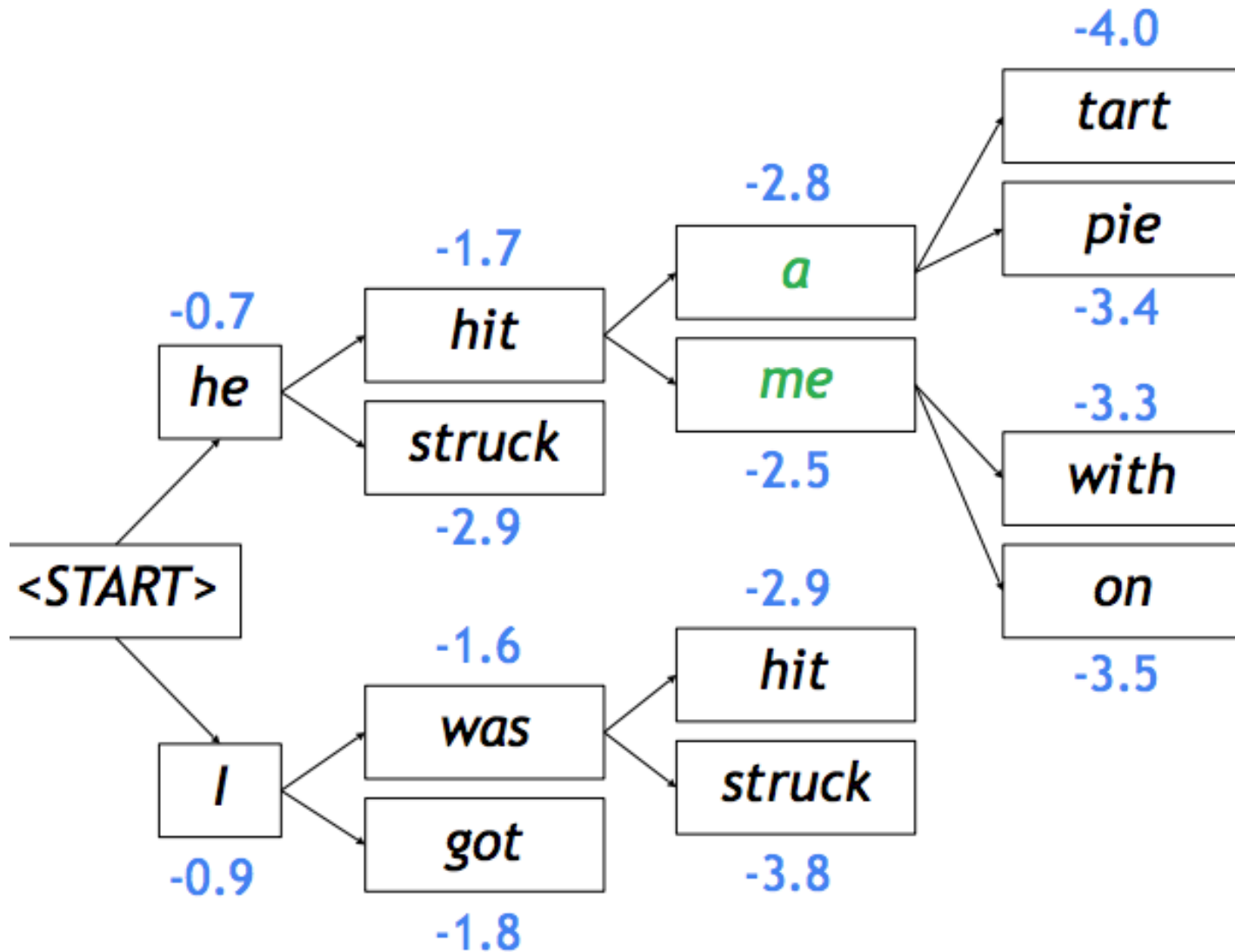
Beam Search: Example



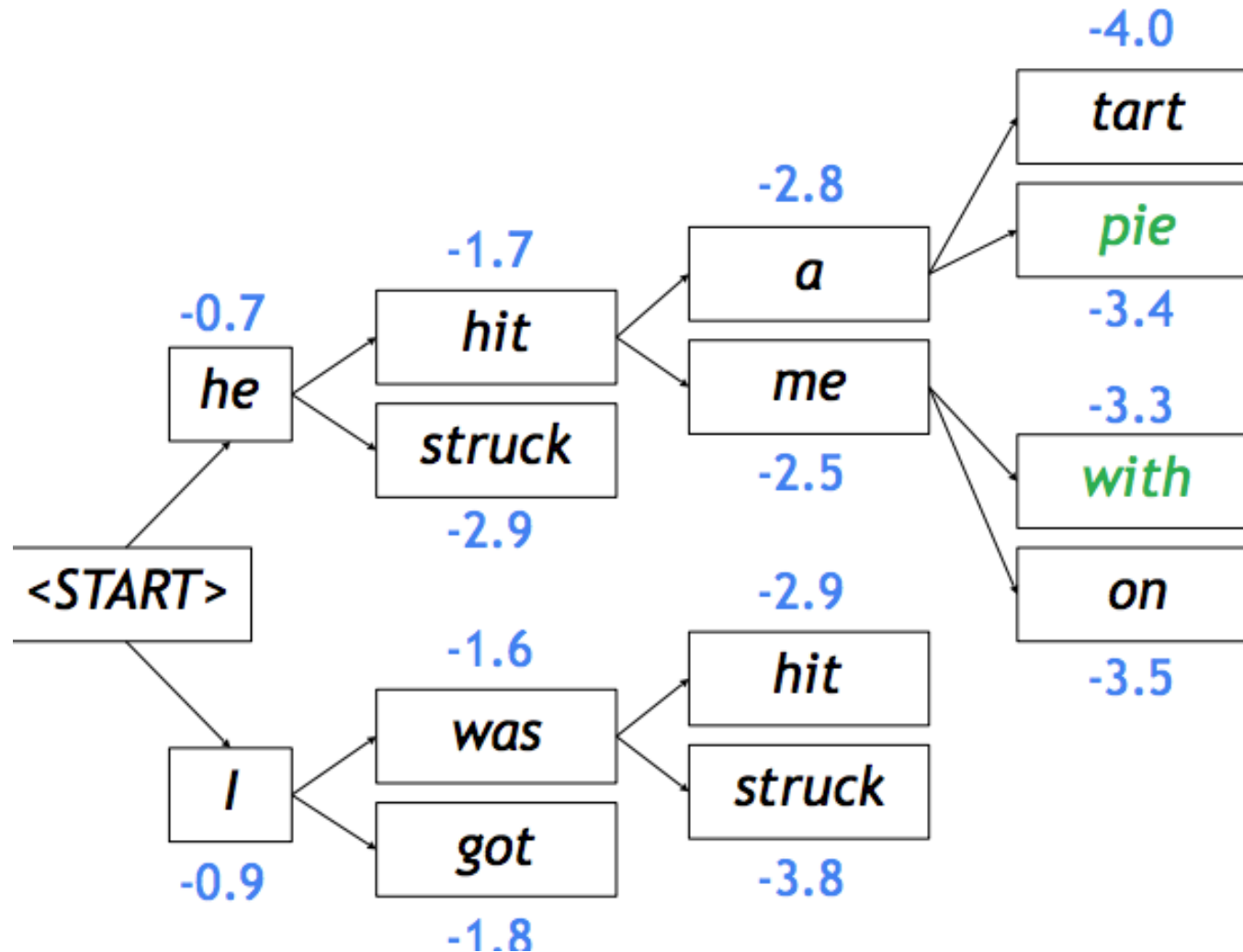
Beam Search: Example



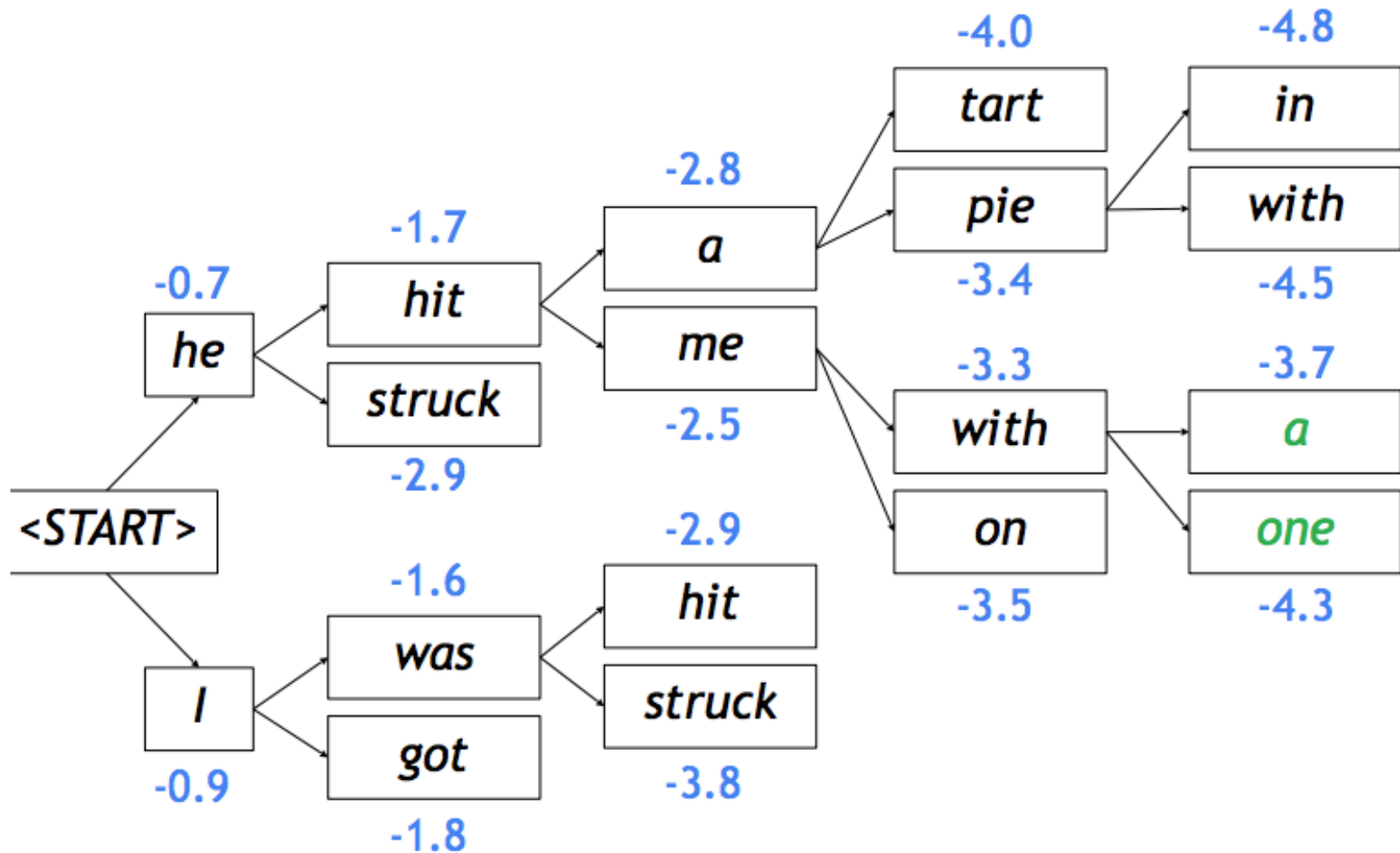
Beam Search: Example



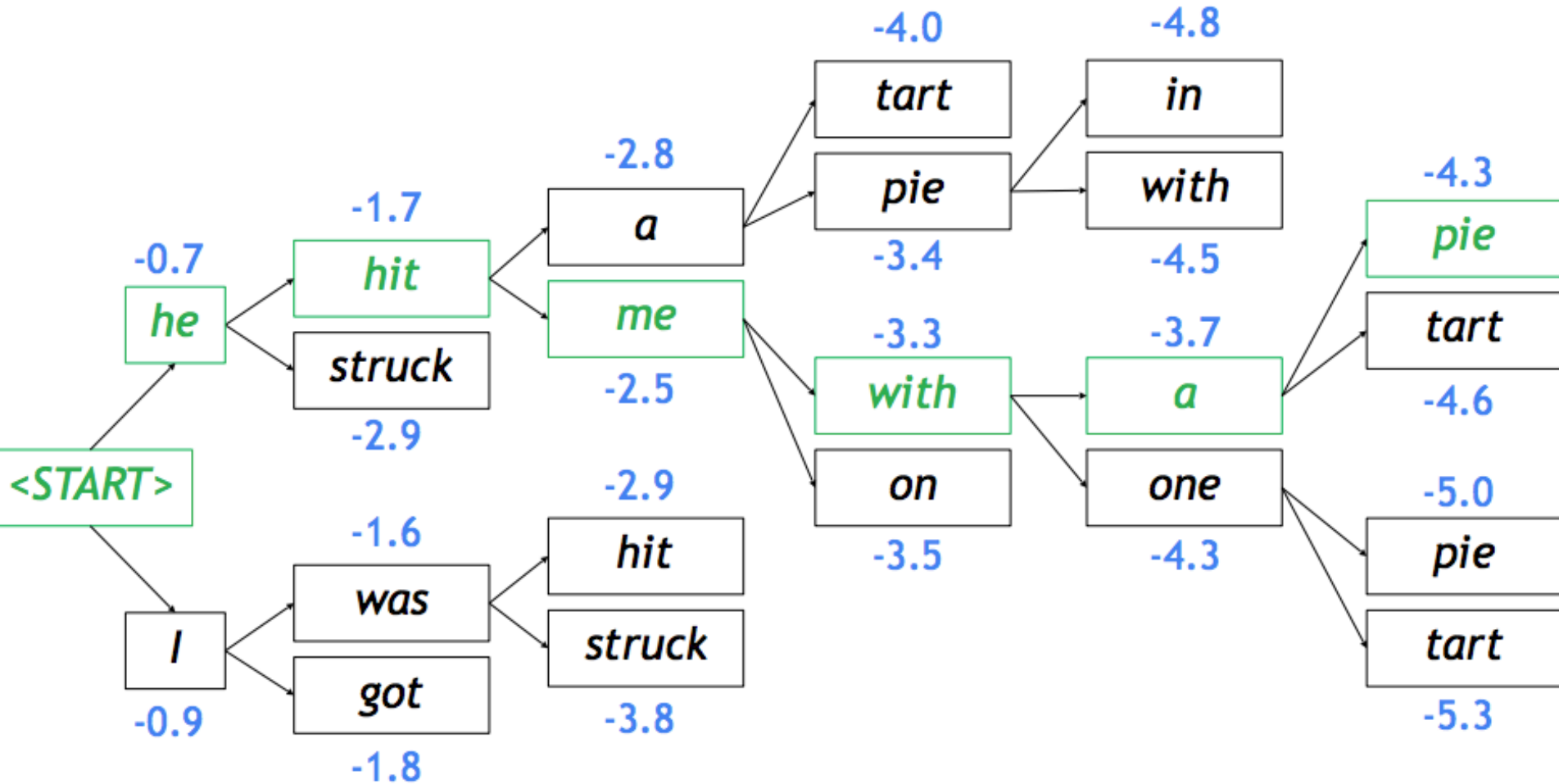
Beam Search: Example



Beam Search: Example



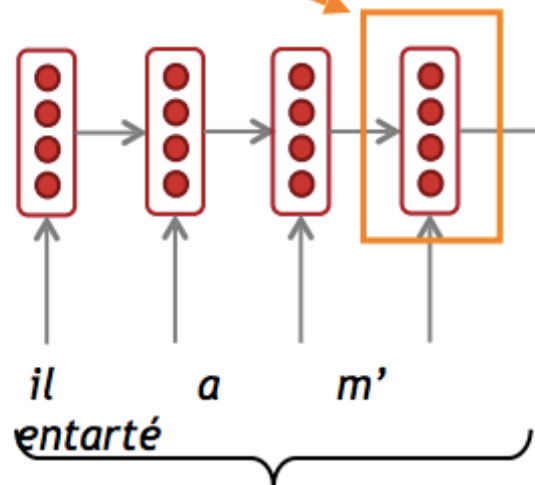
Beam Search: Example



Limitação do Sequence-to-Sequence

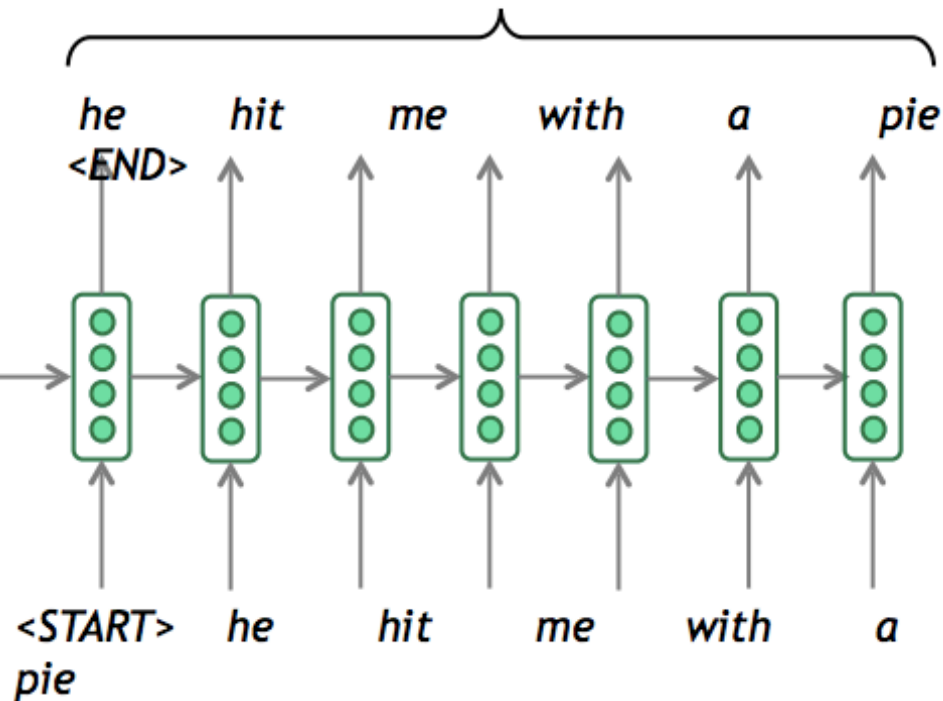
Encoding of the
 source sentence.
 This needs to capture *all*
information about the
 source sentence.
 Information bottleneck!

Encoder RNN



Source sentence (input)

Target sentence (output)

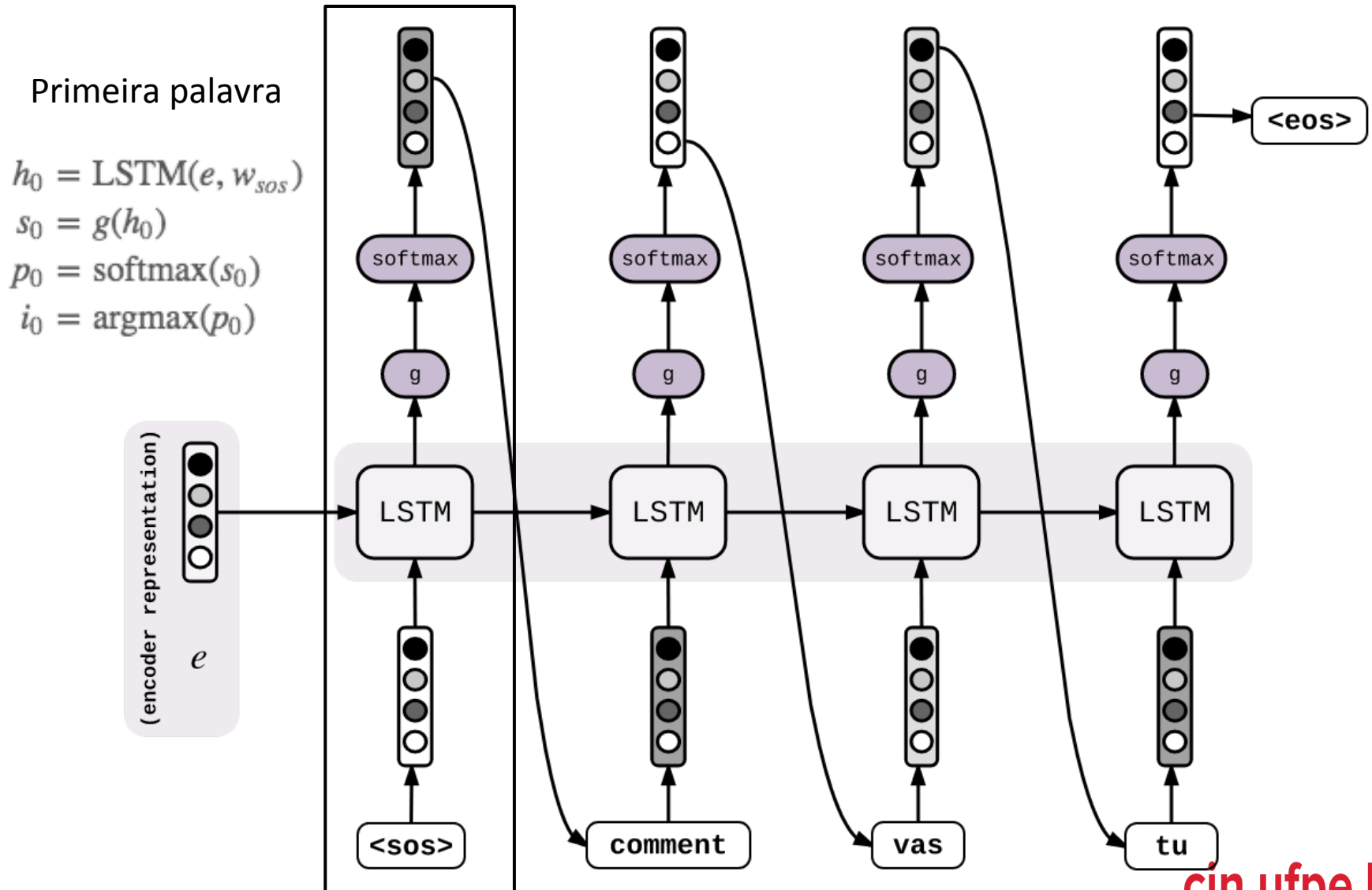


Decoder RNN

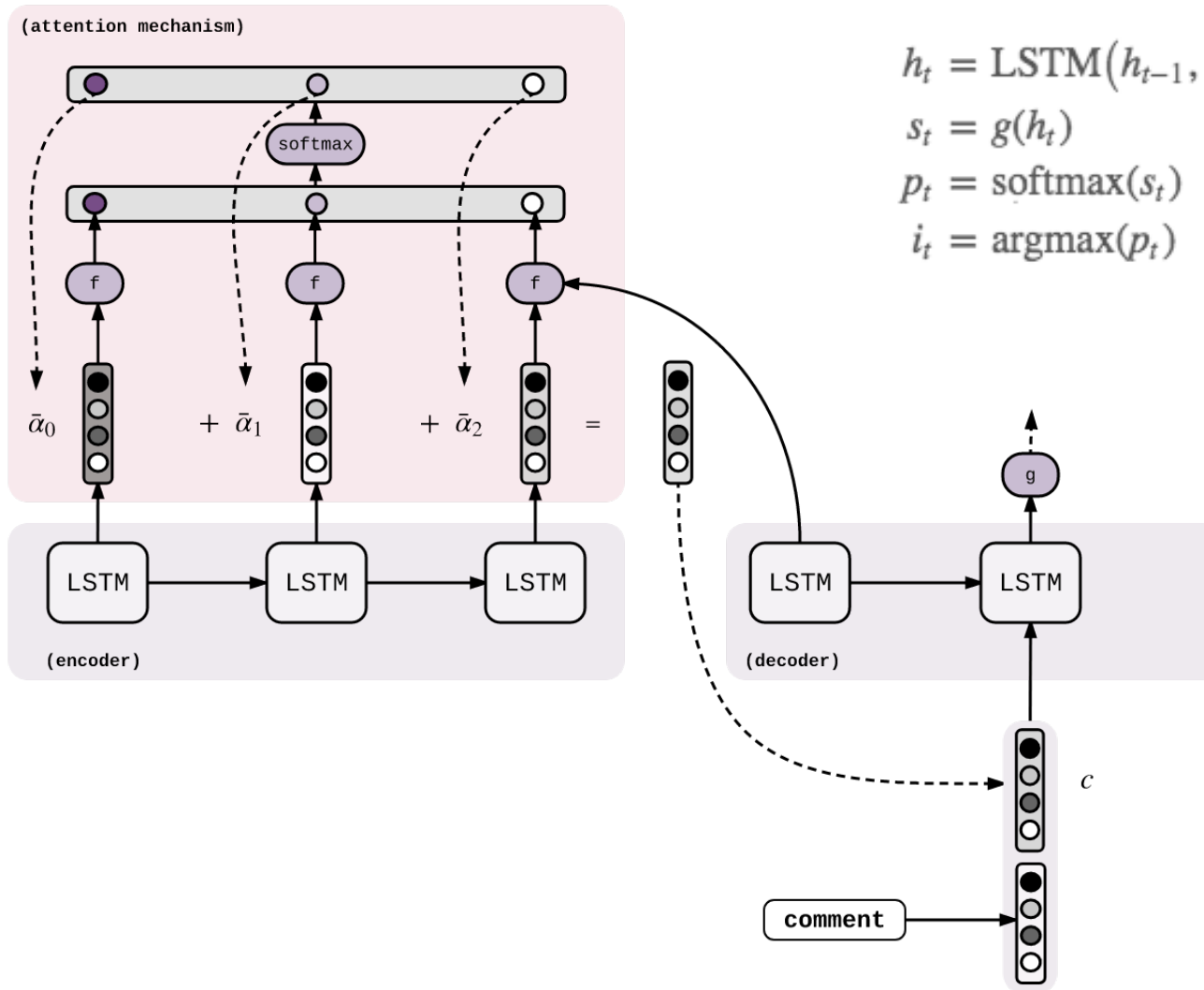
Attention

- Cada passo do decoder tem conexão direta com os encoders para focar em partes específicas da sentença fonte

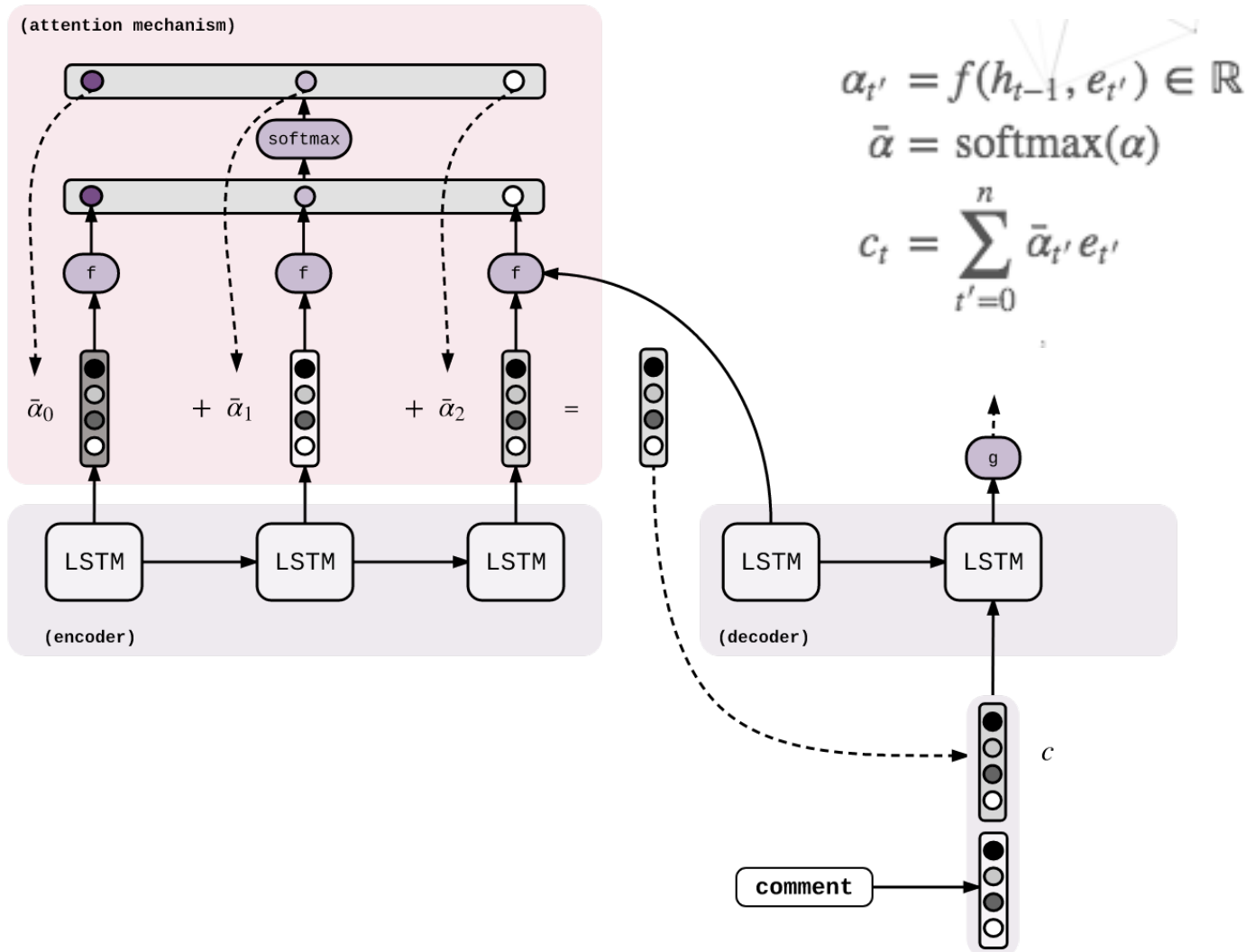
Decoder Tradicional



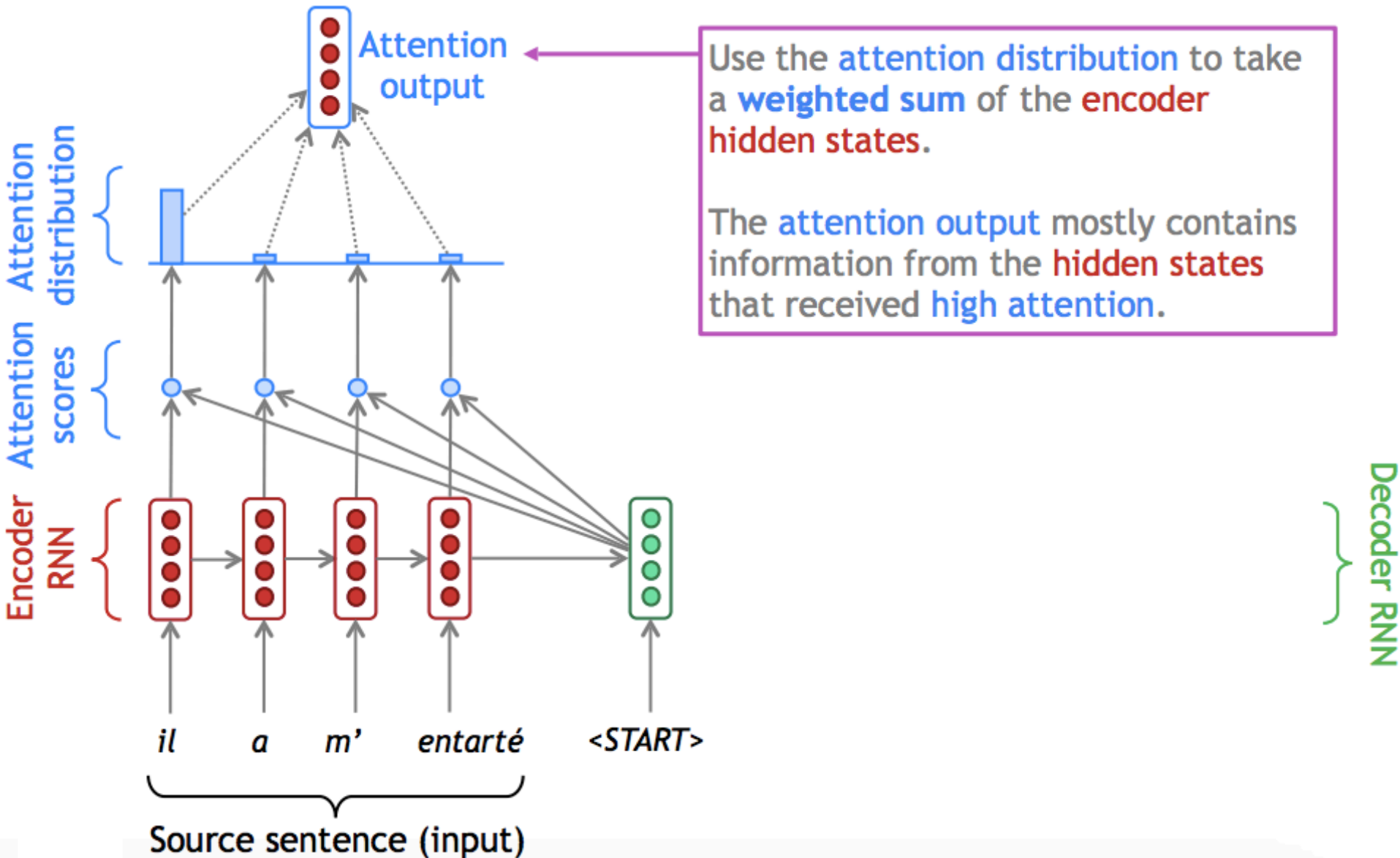
Decoder com Attention



Decoder com Attention



Sequence-to-Sequence com Attention



Funções de Attention

Name	Alignment score function	Citation
Content-base attention	$\text{score}(s_t, h_i) = \text{cosine}[s_t, h_i]$	<u>Graves2014</u>
Additive(*)	$\text{score}(s_t, h_i) = \mathbf{v}_a^\top \tanh(\mathbf{W}_a [s_t; h_i])$	<u>Bahdanau2015</u>
Location-Base	$\alpha_{t,i} = \text{softmax}(\mathbf{W}_a s_t)$ Note: This simplifies the softmax alignment to only depend on the target position.	<u>Luong2015</u>
General	$\text{score}(s_t, h_i) = s_t^\top \mathbf{W}_a h_i$ where \mathbf{W}_a is a trainable weight matrix in the attention layer.	<u>Luong2015</u>
Dot-Product	$\text{score}(s_t, h_i) = s_t^\top h_i$	<u>Luong2015</u>
Scaled Dot-Product(^)	$\text{score}(s_t, h_i) = \frac{s_t^\top h_i}{\sqrt{n}}$ Note: very similar to the dot-product attention except for a scaling factor; where n is the dimension of the source hidden state.	<u>Vaswani2017</u>

Vantagens do Uso de Attention

- Melhora bastante o desempenho do NMT
 - Foca em partes importantes da sentença fonte
- Lida com o problema de vanishing gradient
- Provê alguma interpretabilidade
- Alinhamento das palavras é aprendido automaticamente

	he	hit	me	wit	a	pie
				h		
il						
a						
m'						
entarté						

NMT vs SMT

- Vantagens de NMT
 - Melhor desempenho
 - Mais fluente
 - Melhor uso do contexto
 - Uma única rede neural a ser otimizada
 - Necessita de muito menos esforço:
 - Sem feature engineering
 - Mesmo método para todos os pares de línguas
- Desvantagens de NMT:
 - Menos interpretável
 - Difícil de controlar pois não possui regras

Avaliação de Machine Translation: BLEU Score

- Similaridade entre a tradução automática com traduções feitas por humanos
 - Média geométrica ponderada da n-gram precision (usualmente 1, 2, 3 e 4-grams)
 - Penalidade por traduções muito pequenas
- Varia entre 0 e 1
- 1 muito raro

$$\text{BLEU} = \text{BP} \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right)$$

$p_n =$

$$\frac{\sum_{C \in \{\text{Candidates}\}} \sum_{n\text{-gram} \in C} \text{Count}_{\text{clip}}(n\text{-gram})}{\sum_{C' \in \{\text{Candidates}\}} \sum_{n\text{-gram}' \in C'} \text{Count}(n\text{-gram}')}.$$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

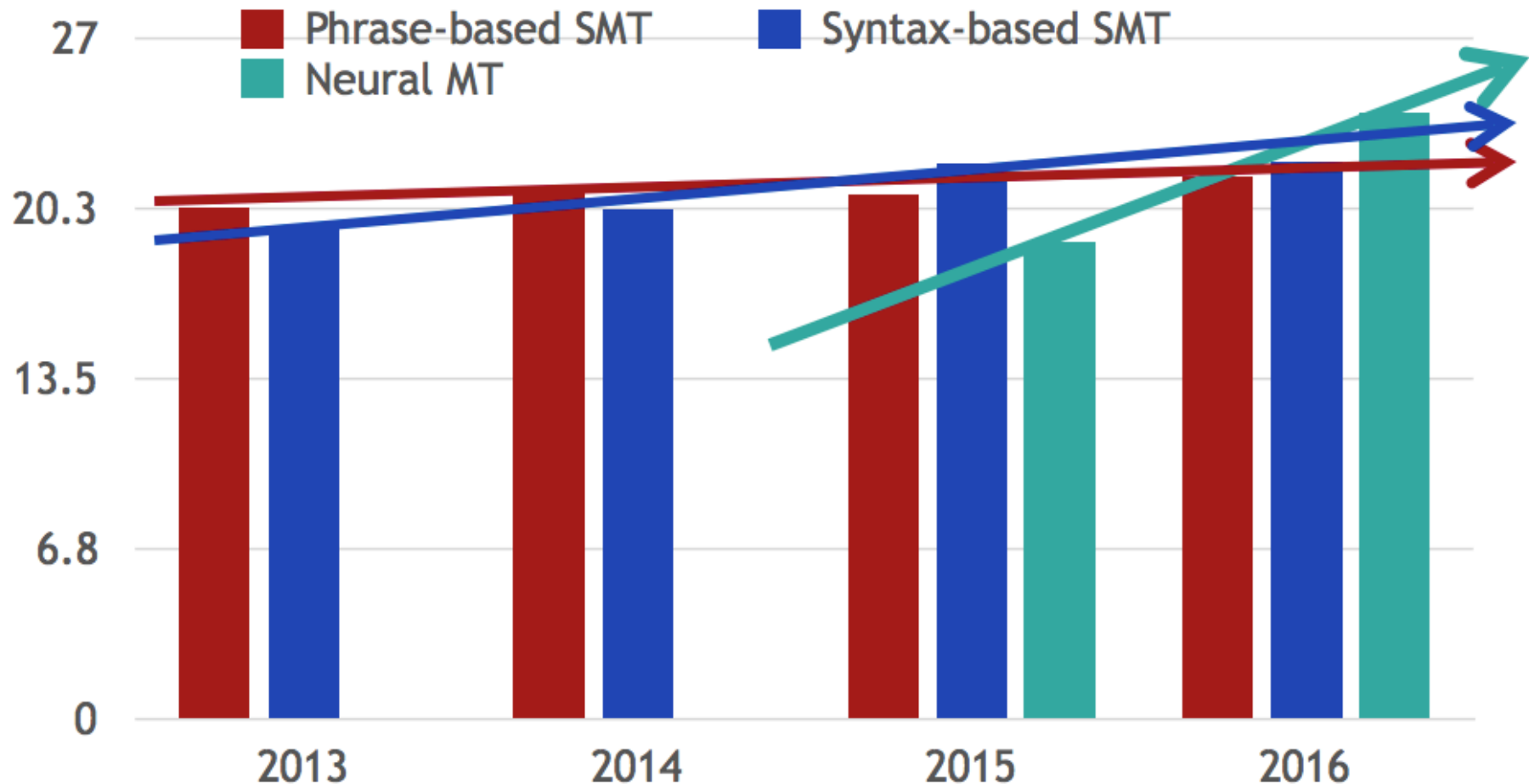
r: referência

c: candidato

BLEU Score: Exemplo

- Referência: “the Iraqi weapons are to be handed over to the army within two weeks”
- MT output: “in two weeks Iraq’s weapons will give army”
- 1-gram precision: 4/8
- 2-gram precision: 1/7
- 3-gram precision: 0/6
- 4-gram precision: 0/5
- BLEU score = 0 (weighted geometric average)

Melhorias dos Modelos com o Tempo



Problemas

- Palavras fora do vocabulário
- Mismatch entre treinamento e teste
- Manter contexto em textos longos
- Línguas com pouco corpus paralelo
- Expressões idiomáticas

