Machine Translation

Luciano Barbosa
(slides baseados no curso de NLP de Stanford)

Centro de Informática UFPE
Universidade Federal de Pernambuco

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*

- Rousseau

Início da Área

- Anos de 1950
- Motivada pela gerra 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

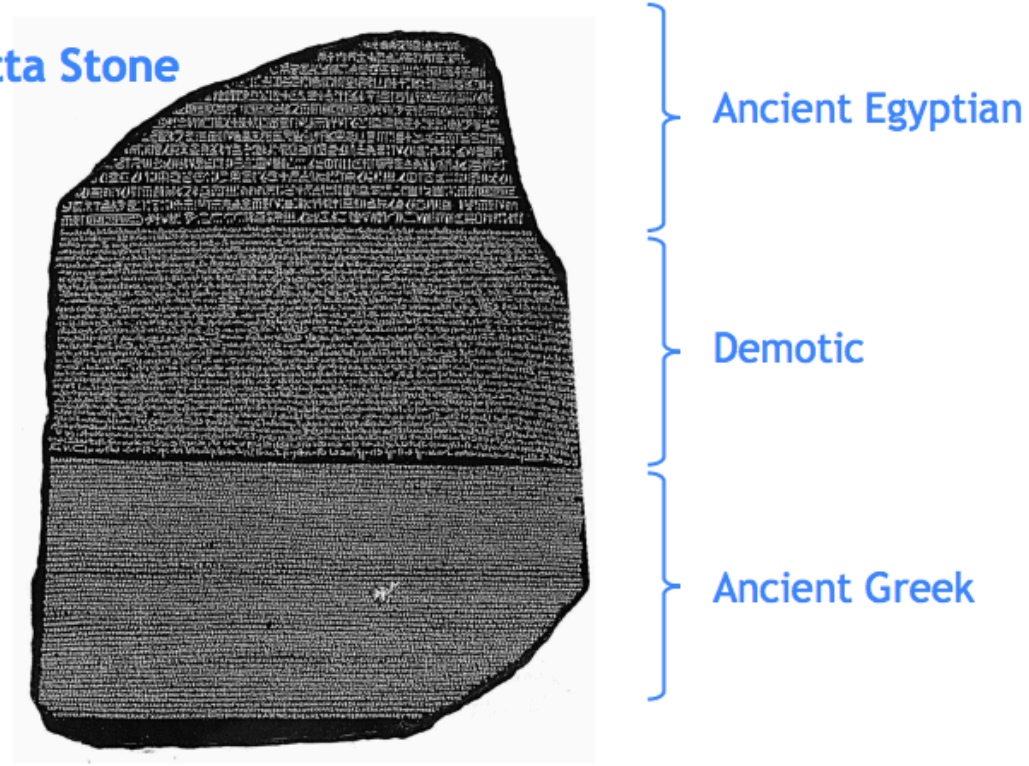
$$\text{argmax}_y P(y|x)$$

- Regra de Bayes quebra em 2 componentes:

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

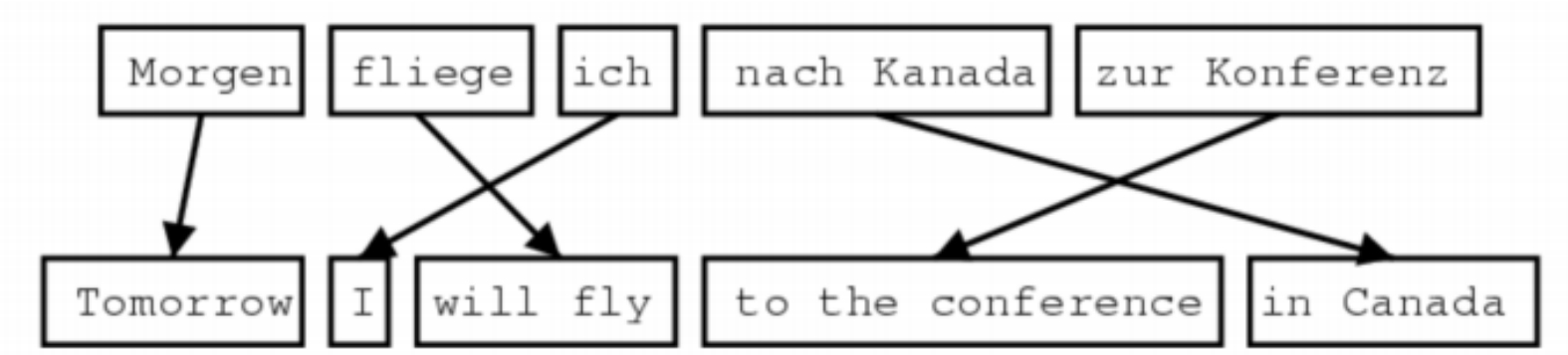
Statistical Machine Translation

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



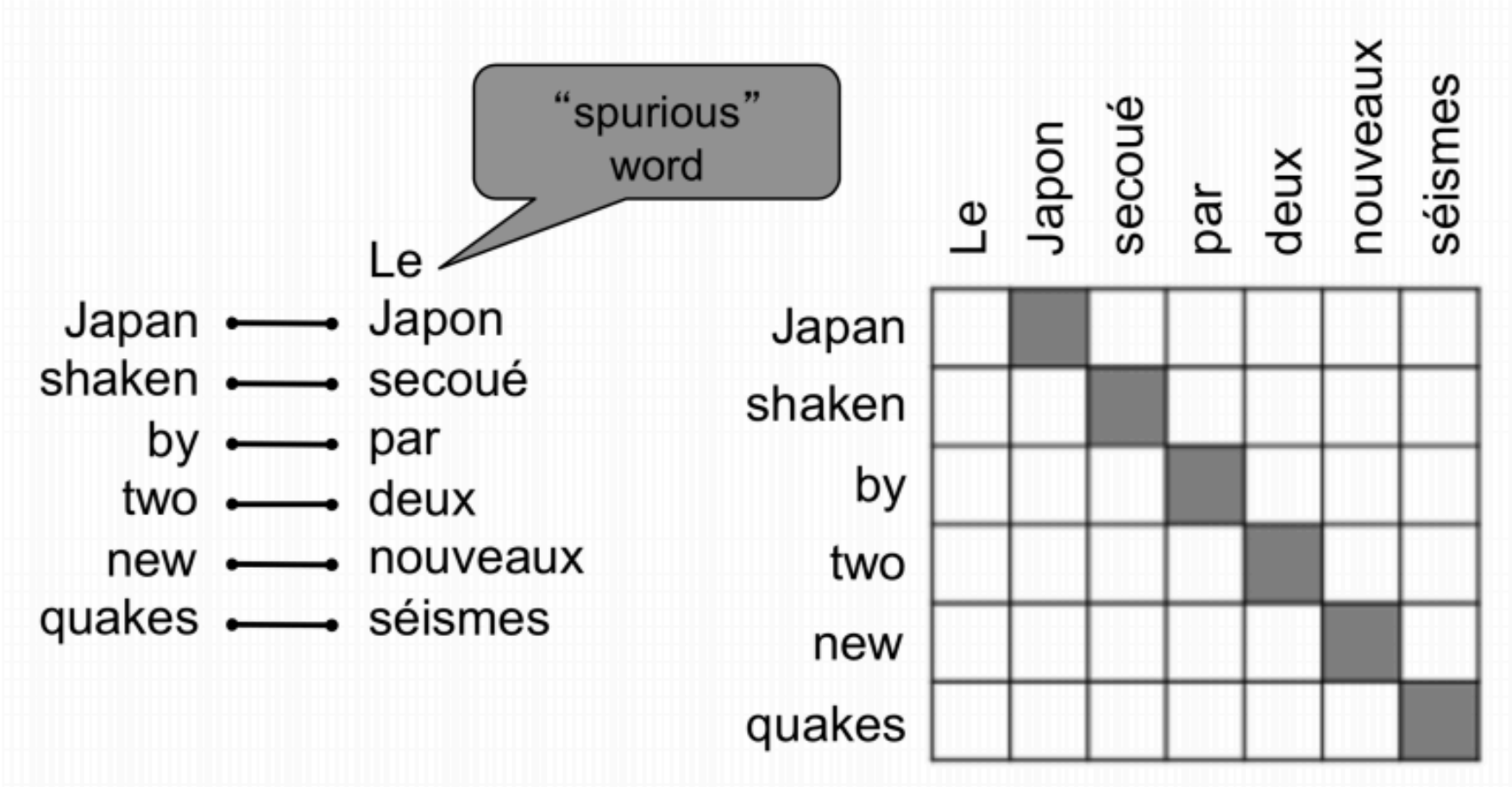
Alinhamento de Corpus Paralelo

- Definir alinhamentos entre sentenças paralelas



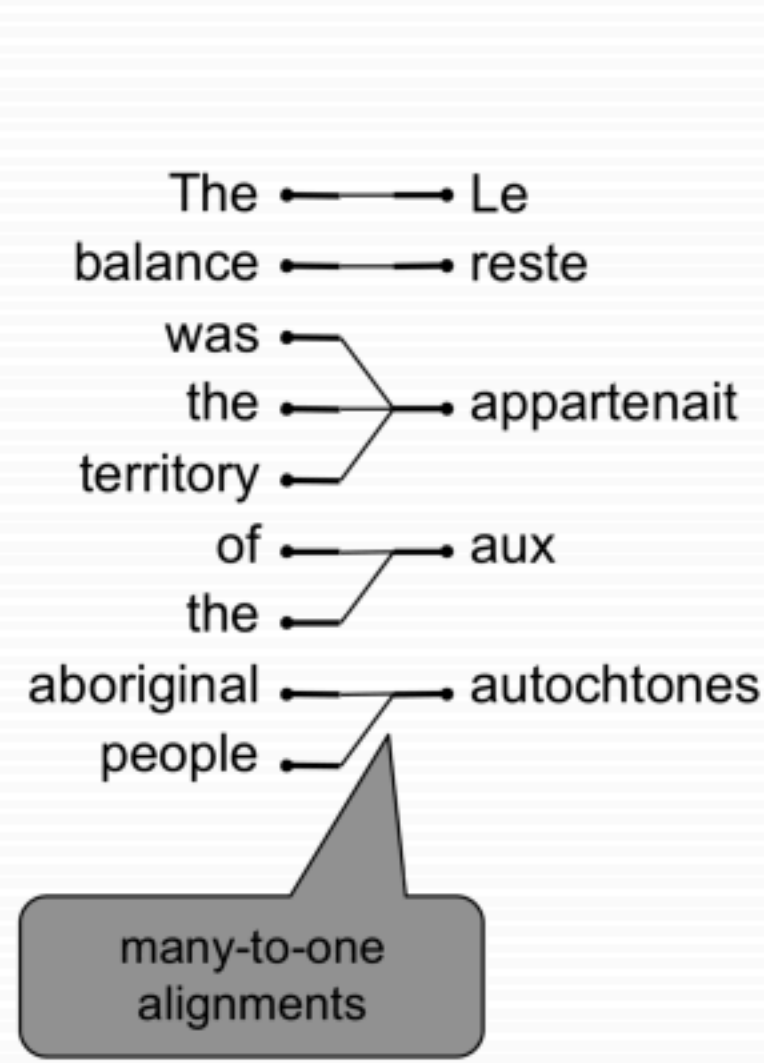
- Desafios:
  - Diferenças na construção de sentenças em diferentes línguas
  - Algumas palavras não possuem tradução

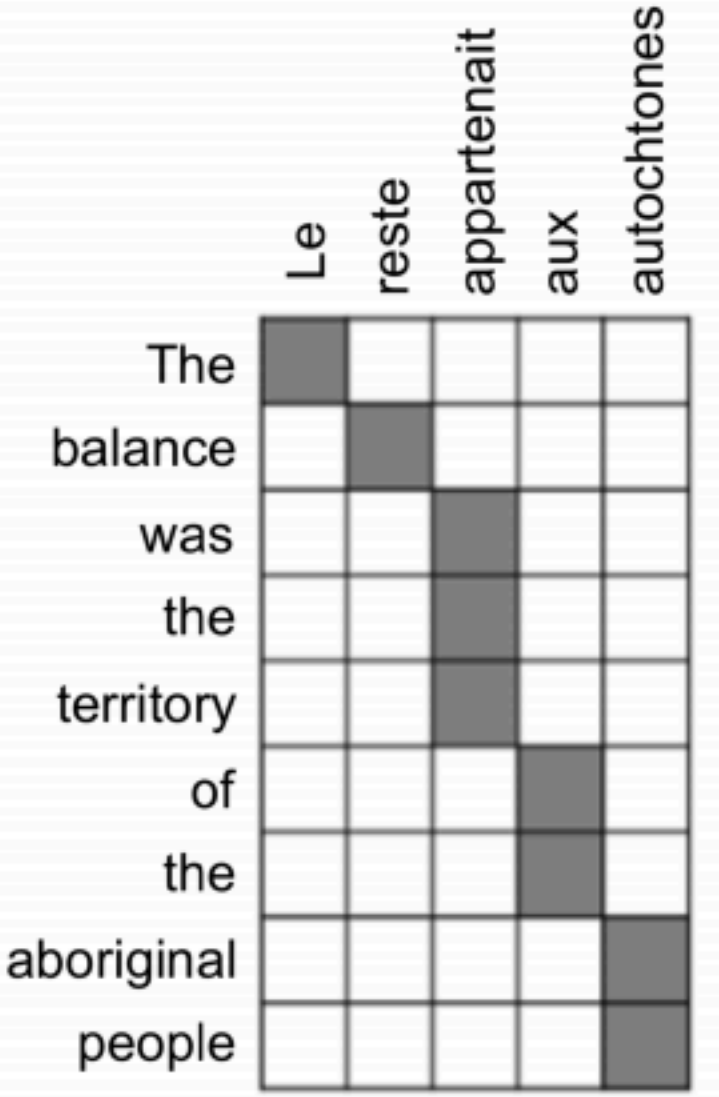
Exemplos



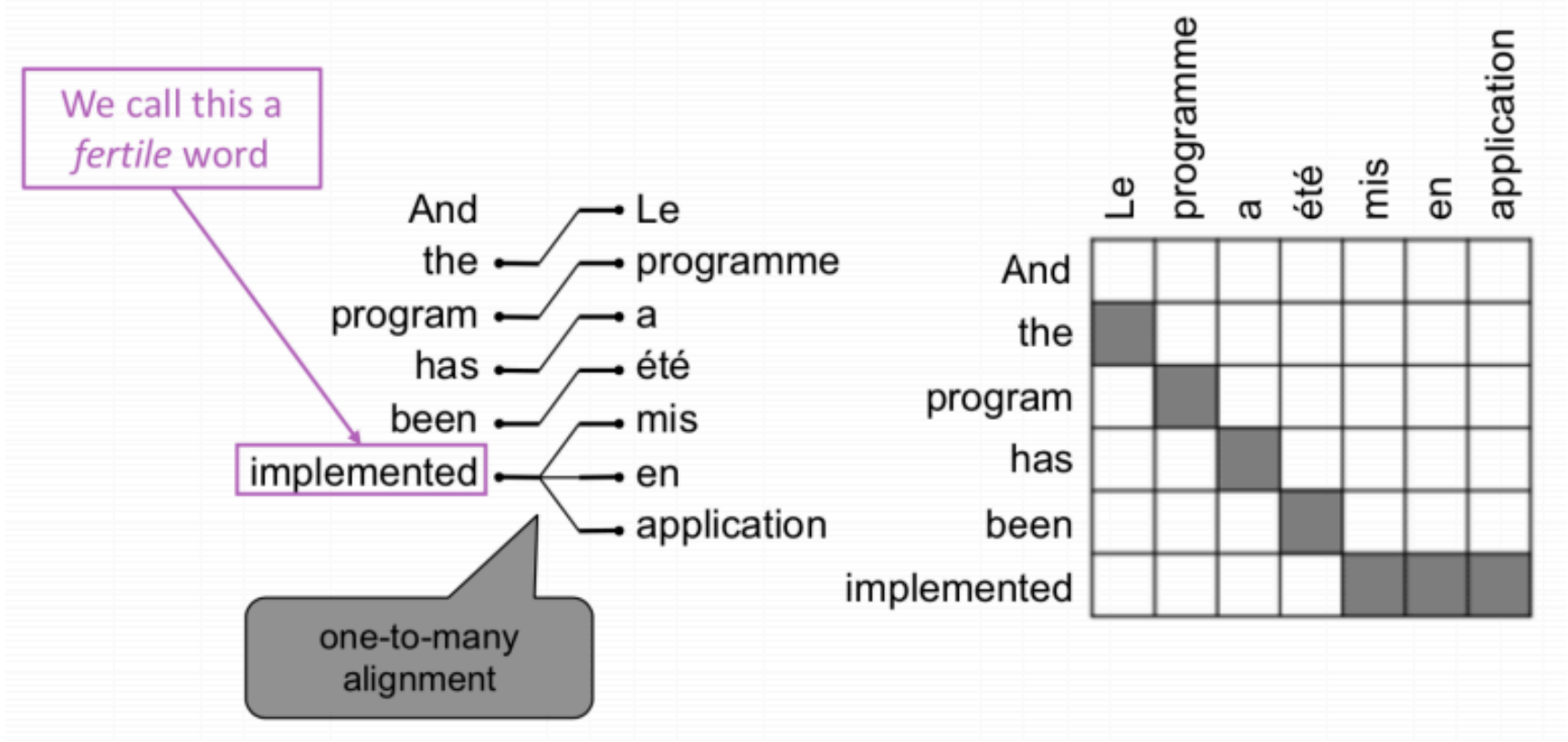
Examples from: "The Mathematics of Statistical Machine Translation: Parameter Estimation", Brown et al, 1993. http://www.aclweb.org/anthology/J93-2003

Exemplos


The
balance
was
the
territory
of
the
aboriginal
people
Le
reste
appartenait
aux
autochtones
many-to-one
alignments

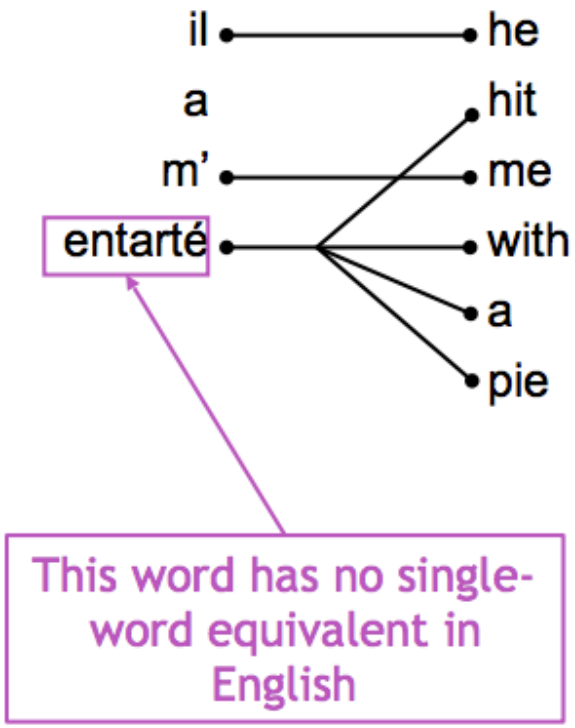

Le
reste
appartenait
aux
autochtones
The
balance
was
the
territory
of
the
aboriginal
people

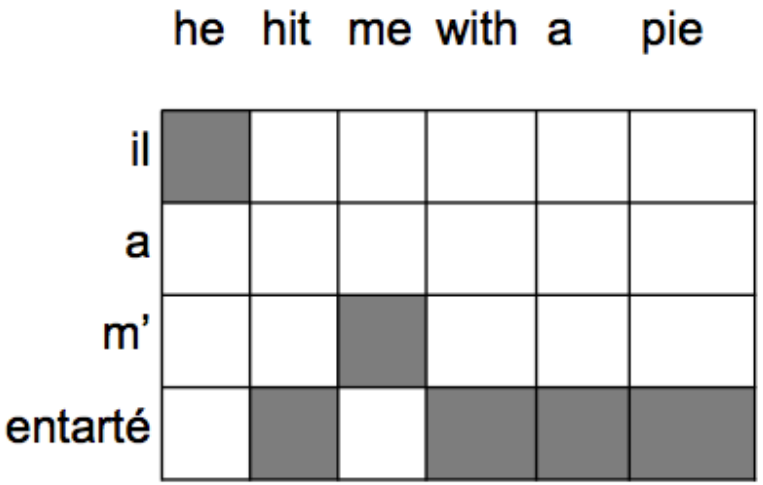
Exemplos



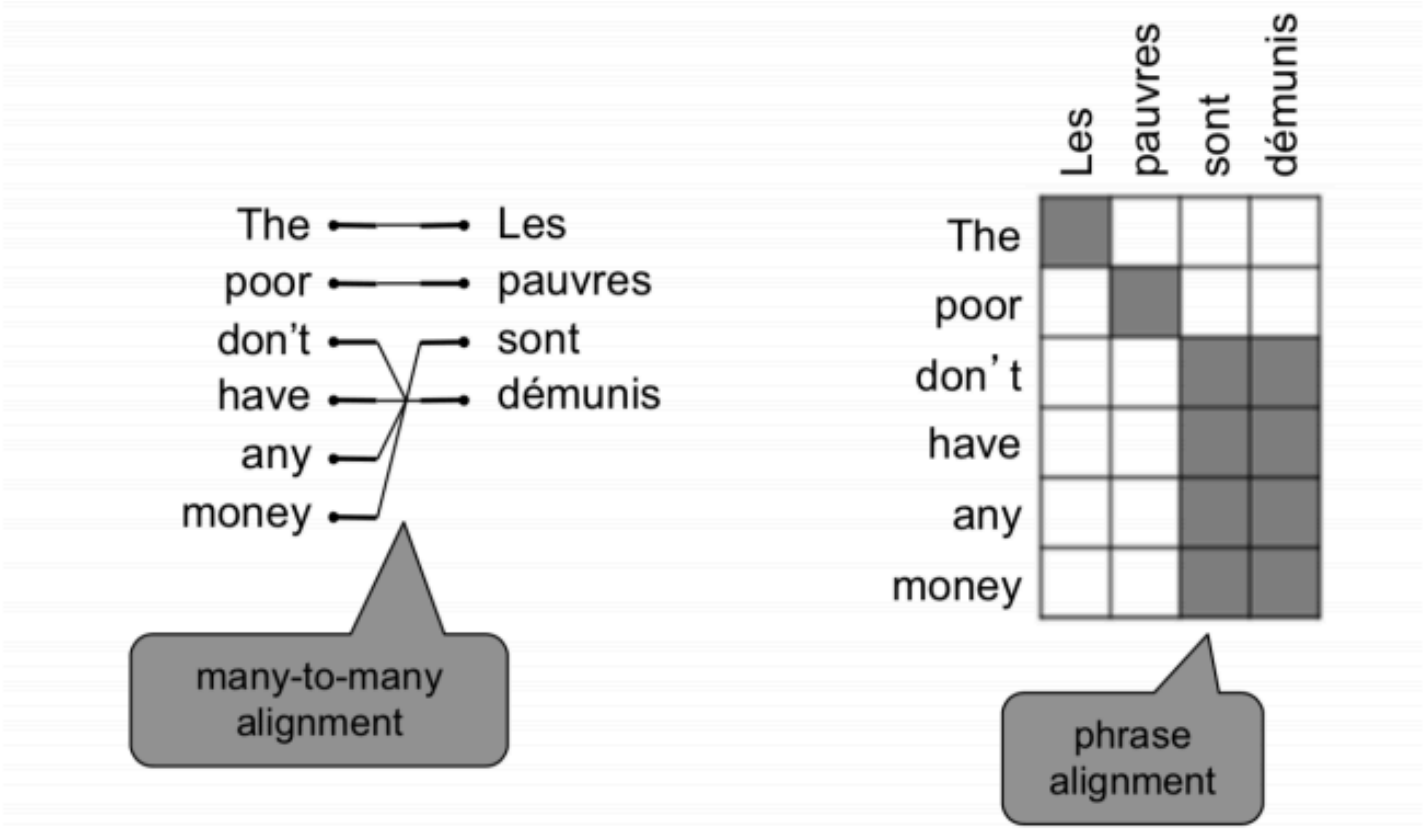
Examples from: "The Mathematics of Statistical Machine Translation: Parameter Estimation", Brown et al, 1993. http://www.aclweb.org/anthology/J93-2003

Exemplos


il
a
m'
entarté
he
hit
me
with
a
pie
This word has no single-word equivalent in English


he hit me with a pie
il
a
m'
entarté

Exemplos



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

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

Question: How to compute this argmax?

- Que hambre tengo yo

argmax_e

What hunger have I	$p(s\|e)p(e) = 0.000014 \times 0.000001$
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	$p(s\|e)p(e) = 0.000020 \times 0.00000098$
...	

= I am so hungry

https://www.slideshare.net/hongjoo/koreanizer-statistical-machine-translation-based-roko-transliterator-164610918

Statistical Machine Translation

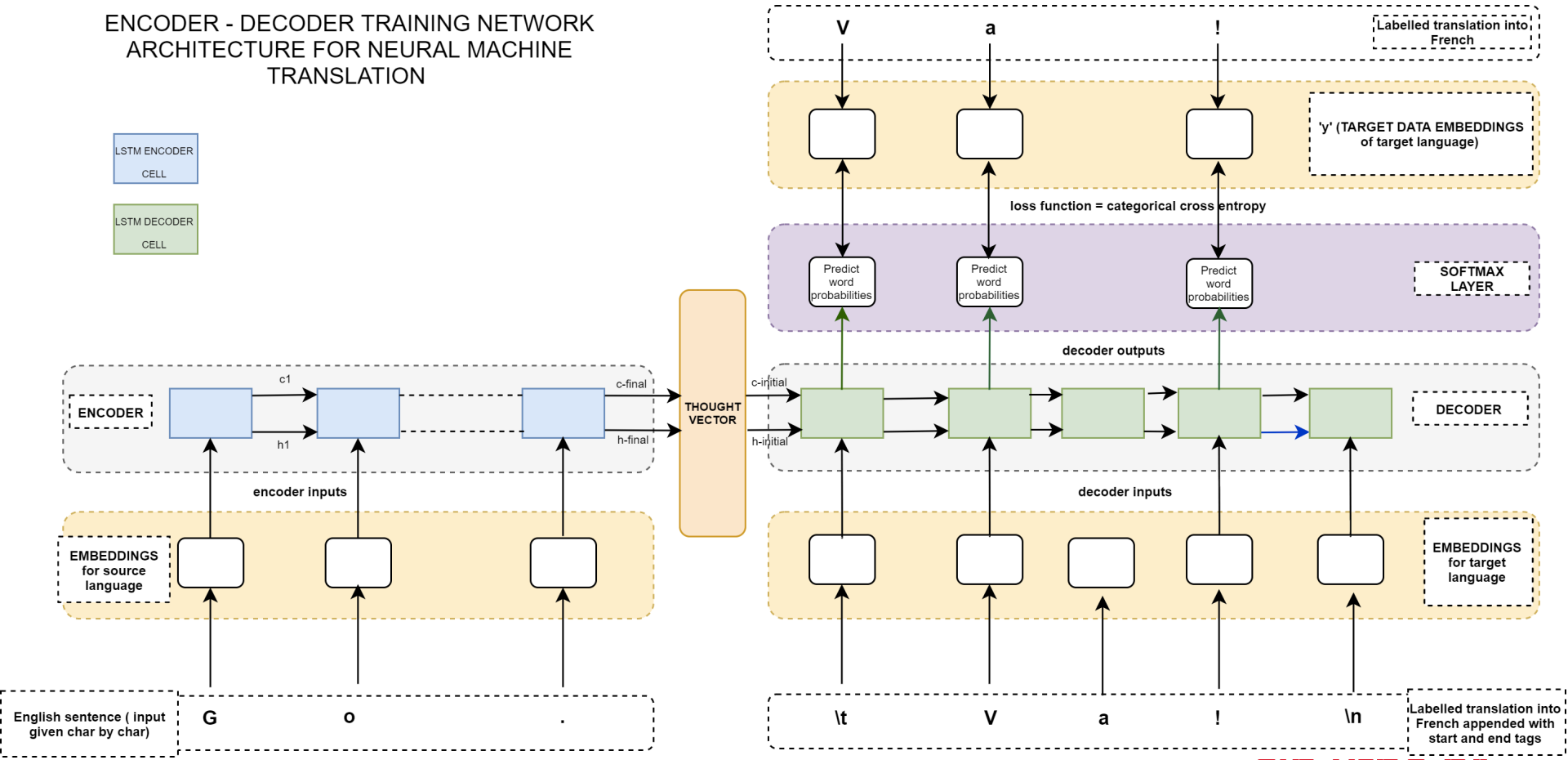
- Melhores sistemas bem complexos
- Feature engineering para capturar detalhes das linguagens
- Muito esforço humano para manter

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

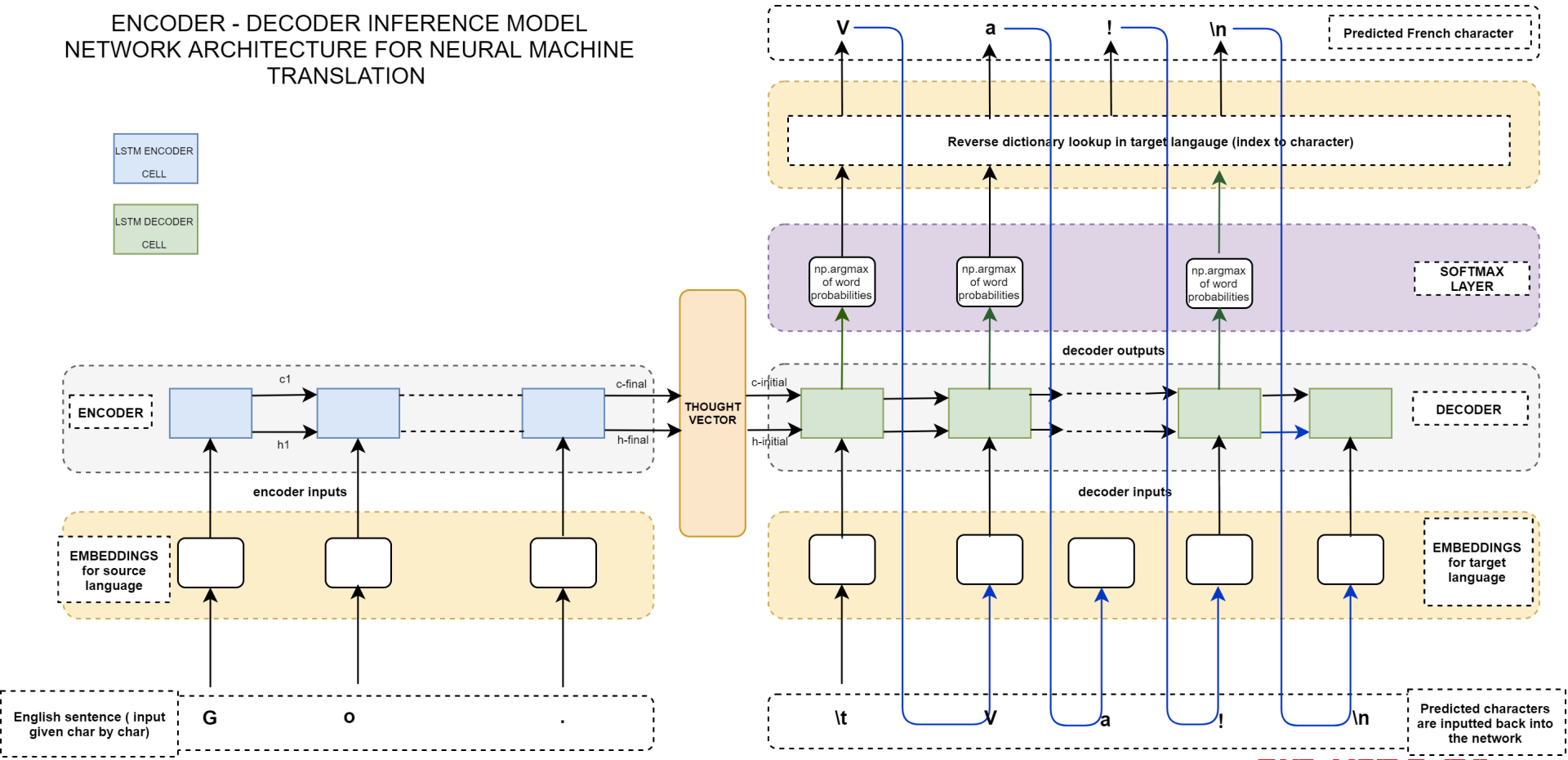
- Precisa de corpus paralelo



https://towardsdatascience.com/neural-machine-translation-using-seq2seq-with-keras-c23540453c74

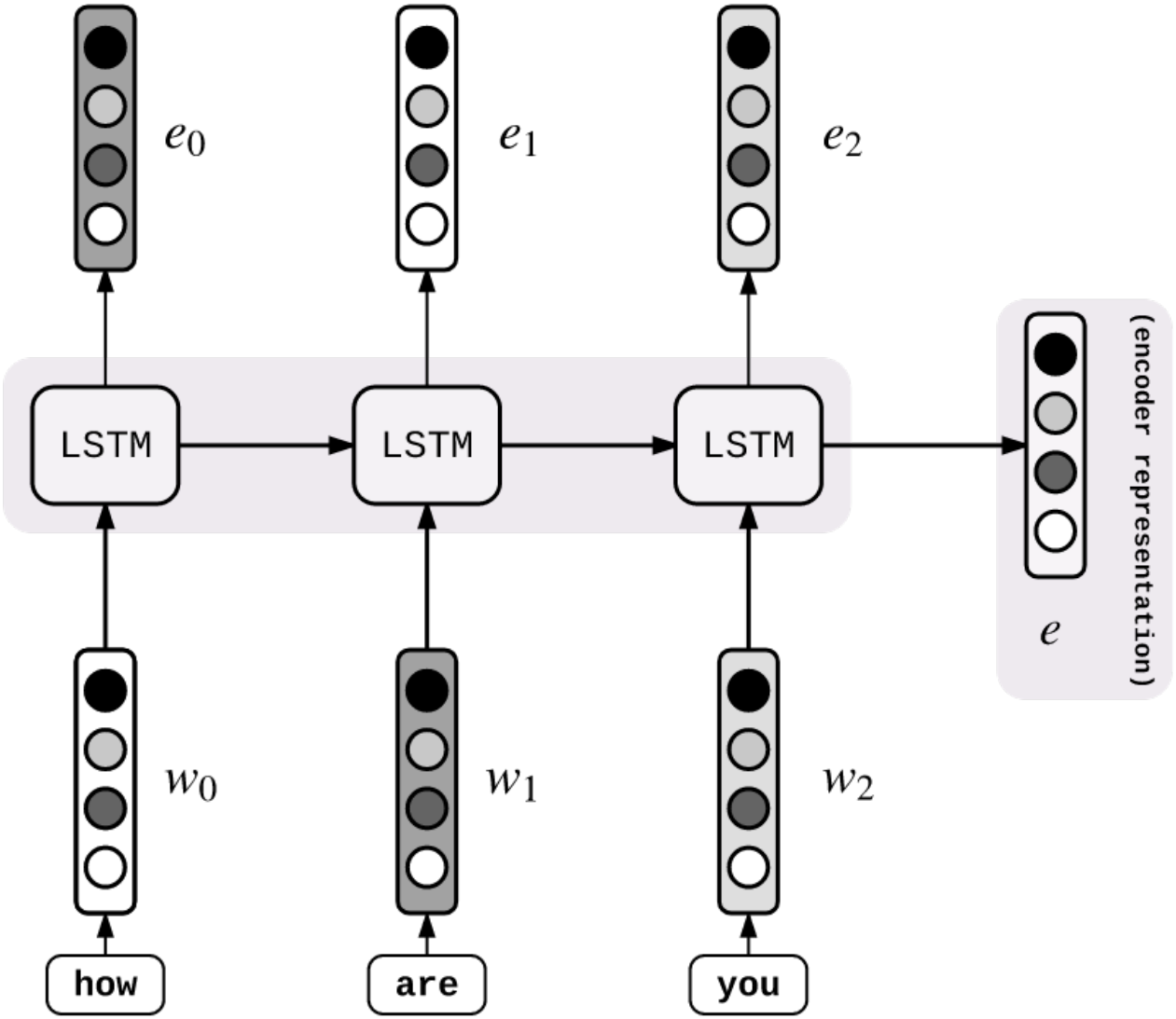
Tradução

- Greedy: Seleciona a palavra com máxima probabilidade



https://towardsdatascience.com/neural-machine-translation-using-seq2seq-with-keras-c23540453c74

Encoder

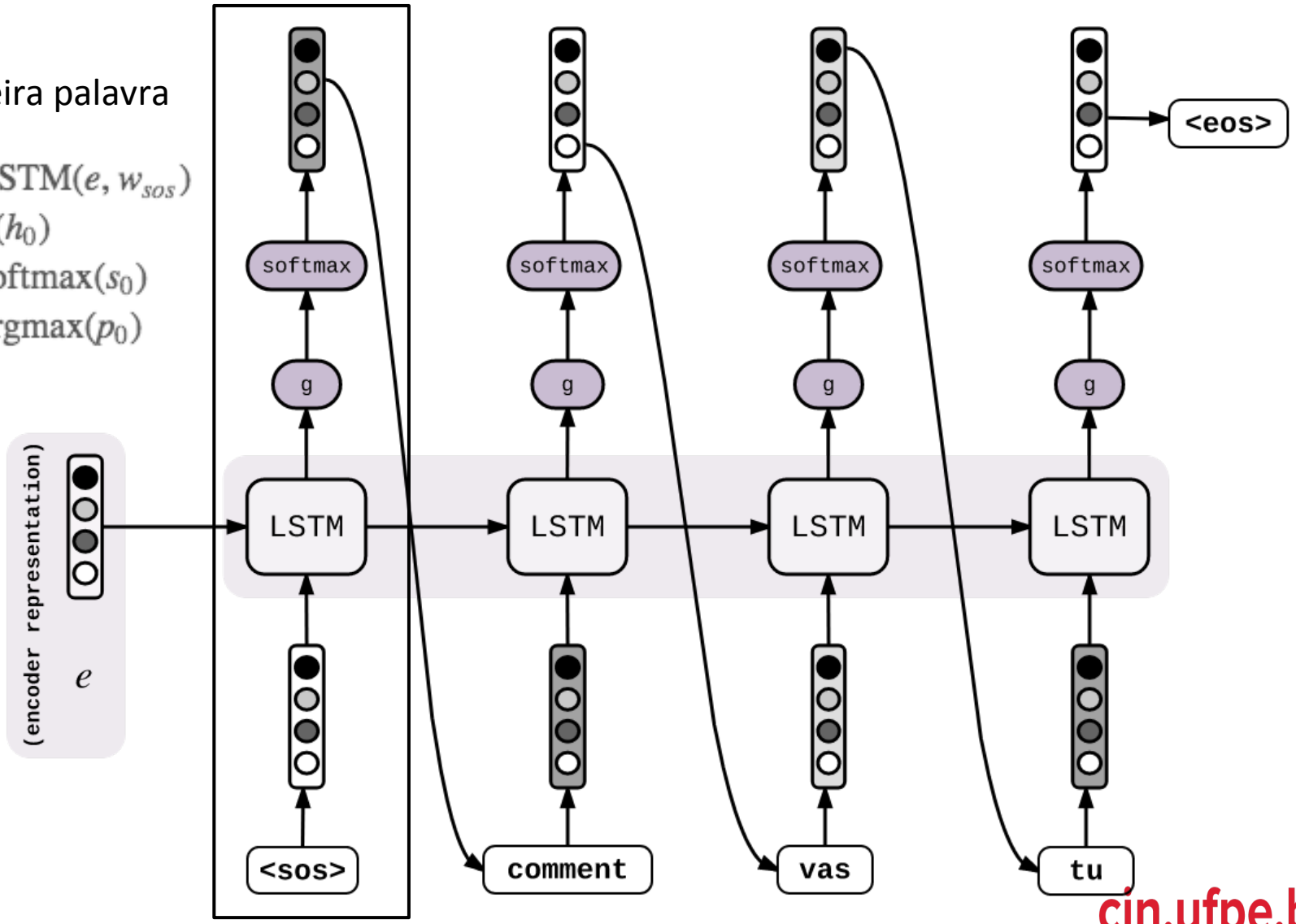


https://tomaxent.com/2018/01/21/Seq2Seq-with-Attention-and-Beam-Search-Repost/

Decoder (Greedy)

Primeira palavra

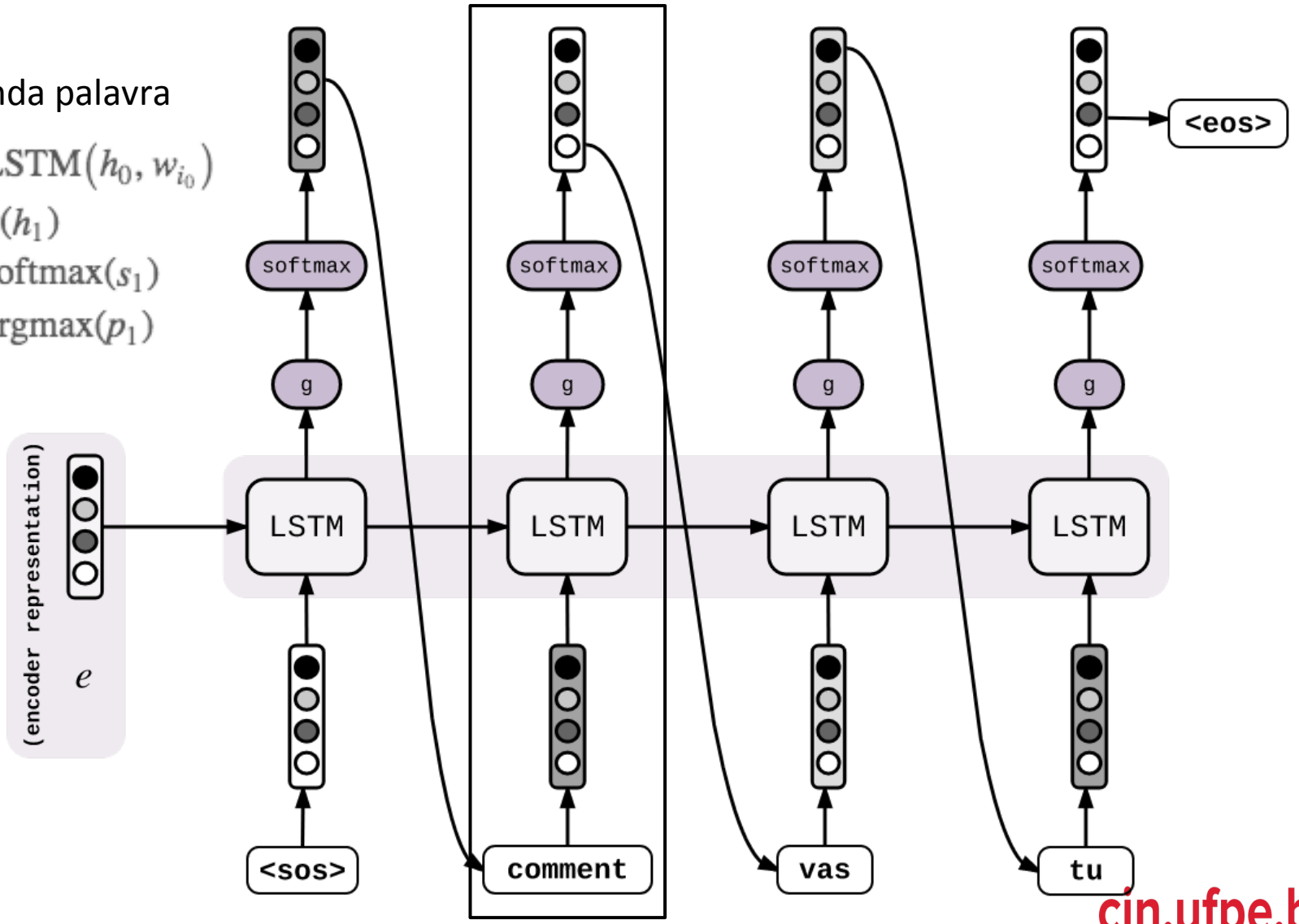
$$h_0 = \text{LSTM}(e, w_{sos})$$
$$s_0 = g(h_0)$$
$$p_0 = \text{softmax}(s_0)$$
$$i_0 = \text{argmax}(p_0)$$



Decoder (Greedy)

Segunda palavra

$$h_1 = \text{LSTM}(h_0, w_{i_0})$$
$$s_1 = g(h_1)$$
$$p_1 = \text{softmax}(s_1)$$
$$i_1 = \text{argmax}(p_1)$$

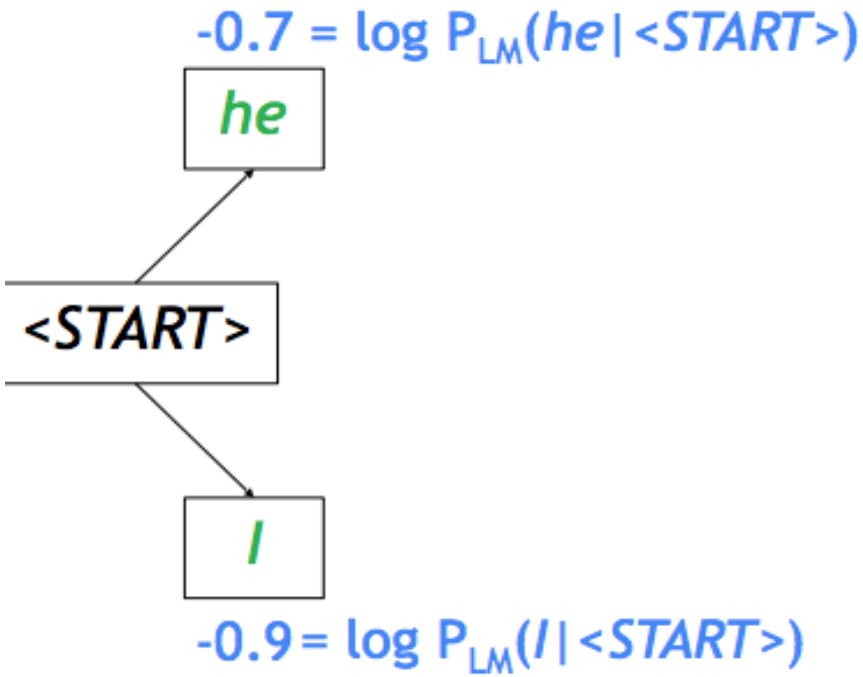


Beam Search

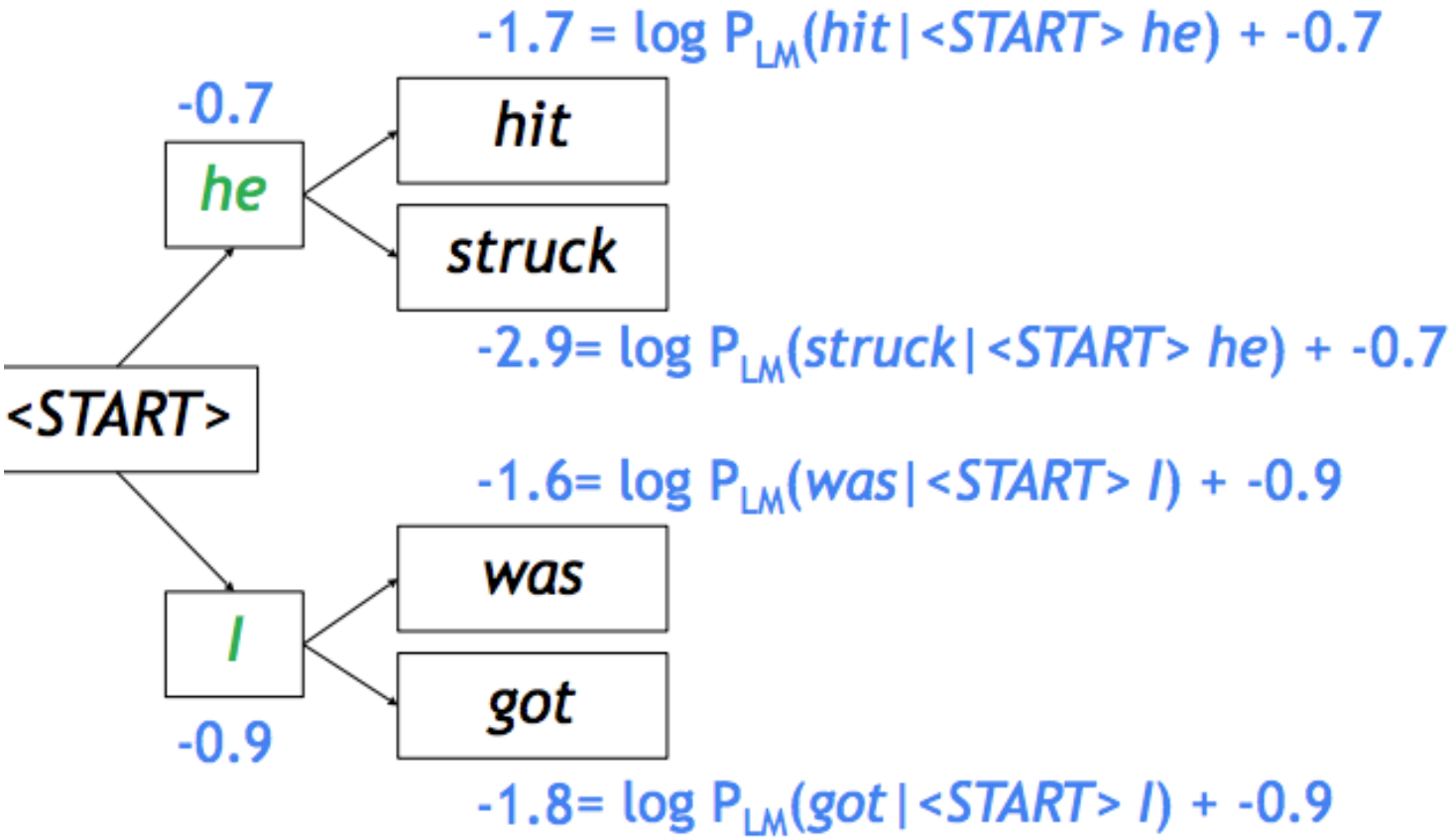
- Limitação do greedy decoding: não permite desfazer decisões
- Em cada passo, manter as k traduções parciais mais prováveis ou hipóteses
- Não é garantido encontrar a melhor solução

Beam Search: Example

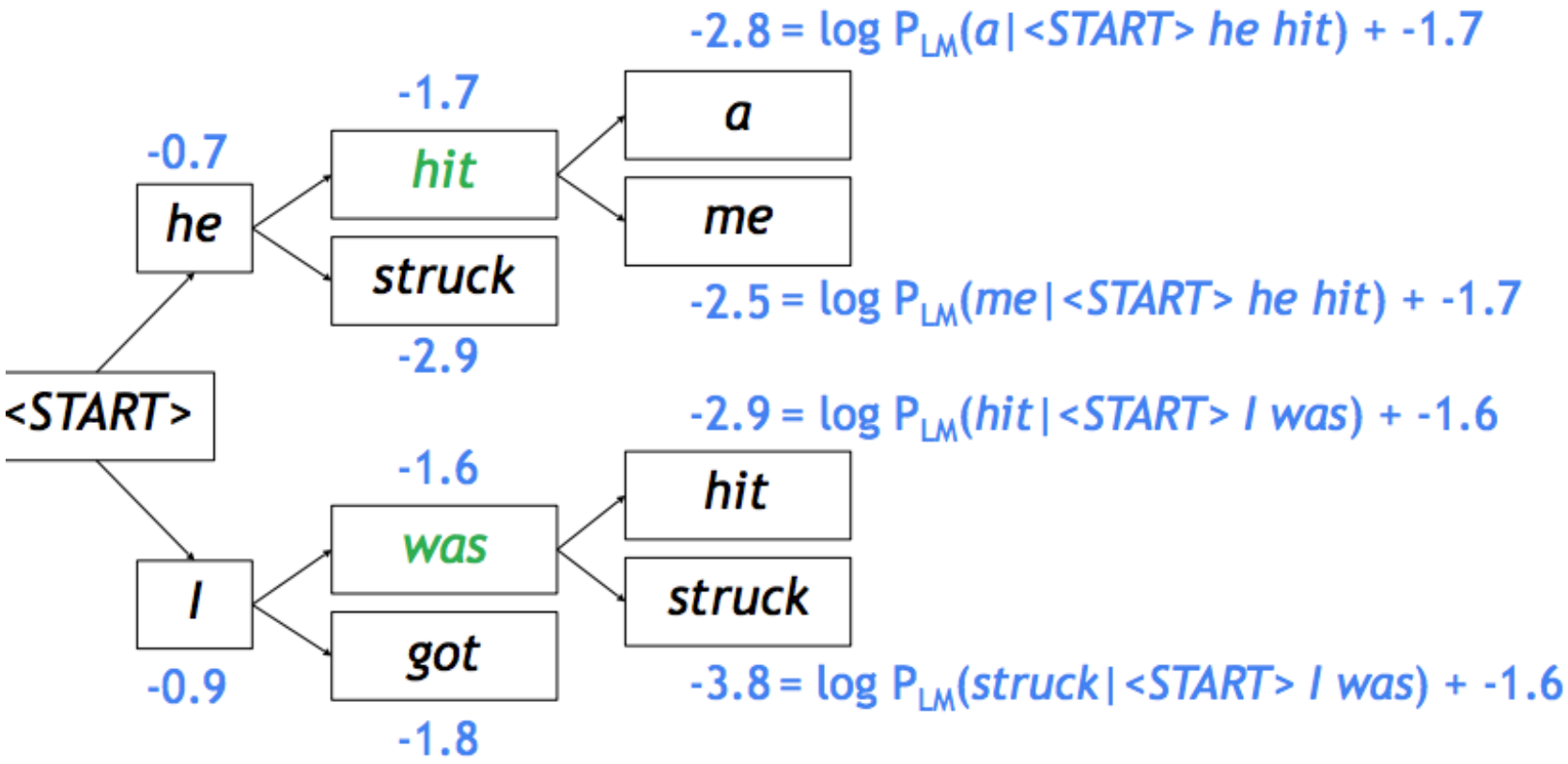
- Source: il a m’ entarté
- Target: he hit me with a pie
- K=2



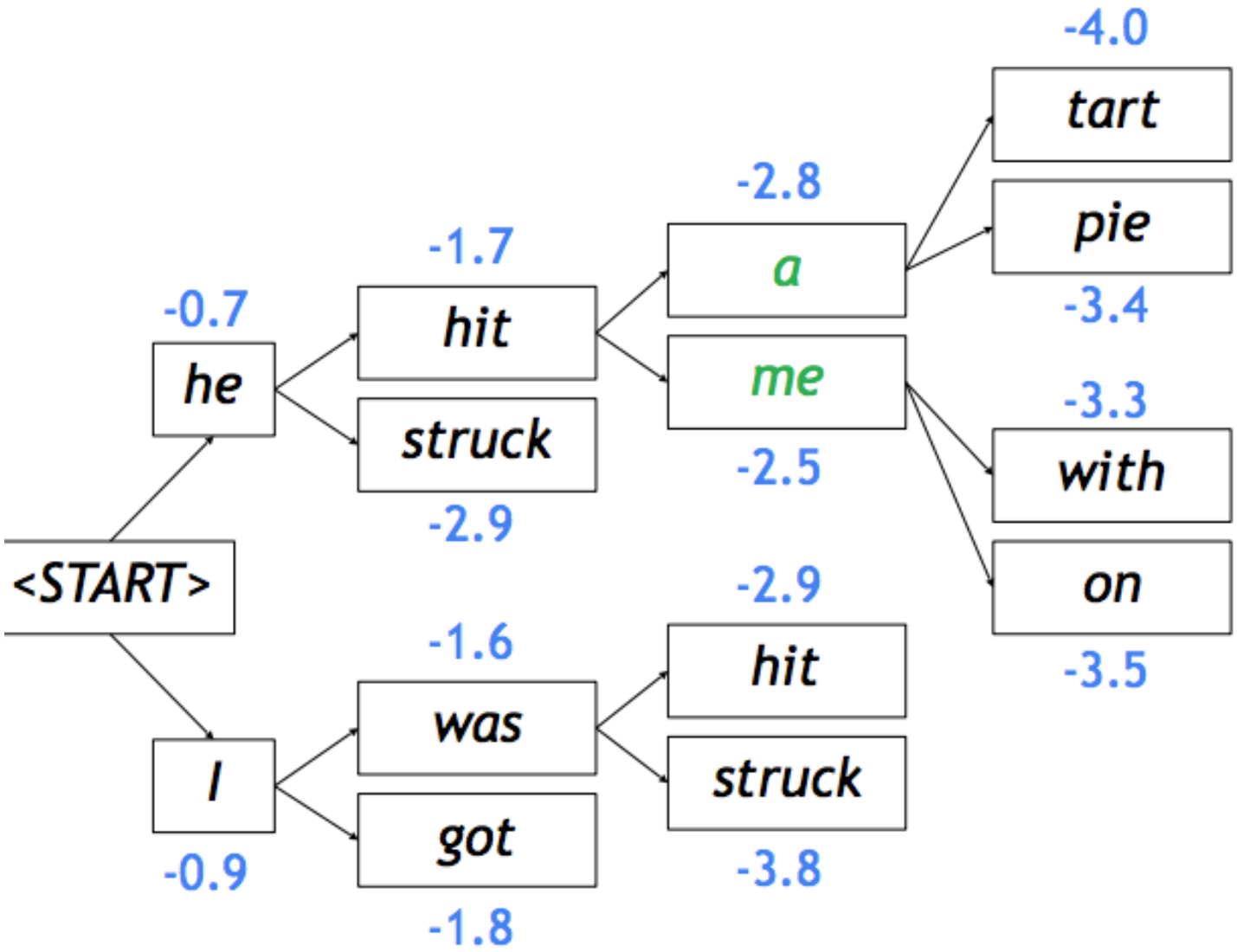
Beam Search: Example



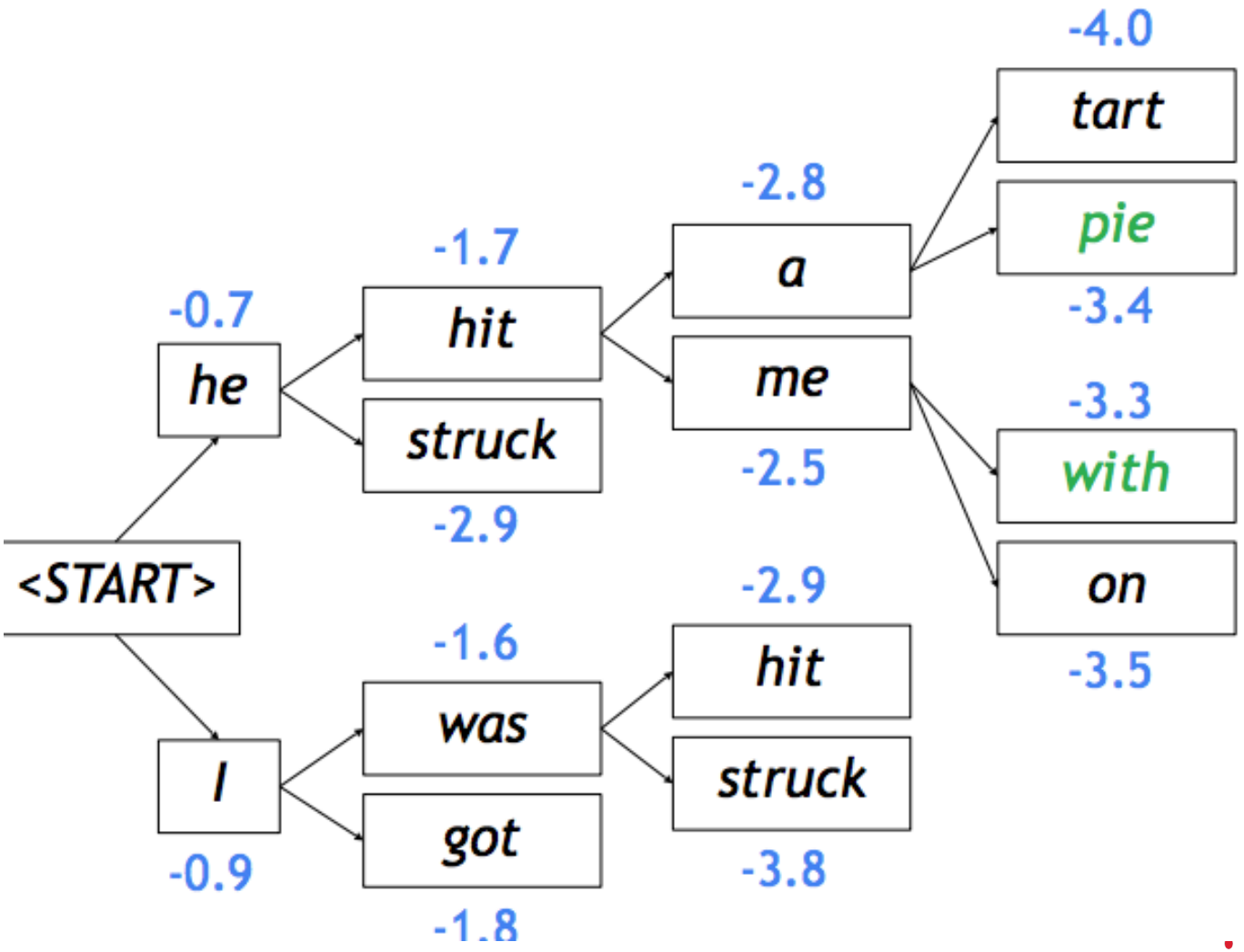
Beam Search: Example



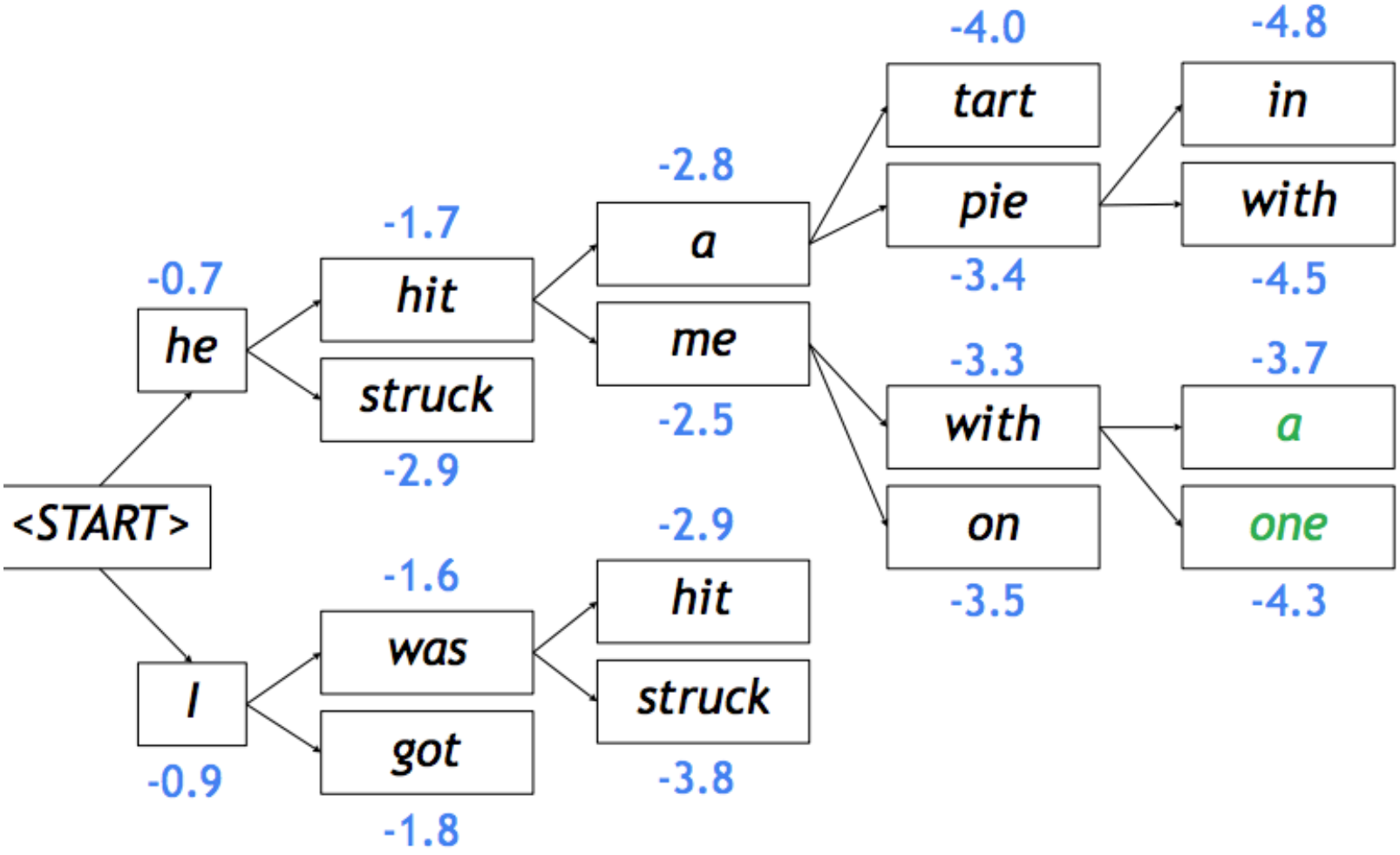
Beam Search: Example



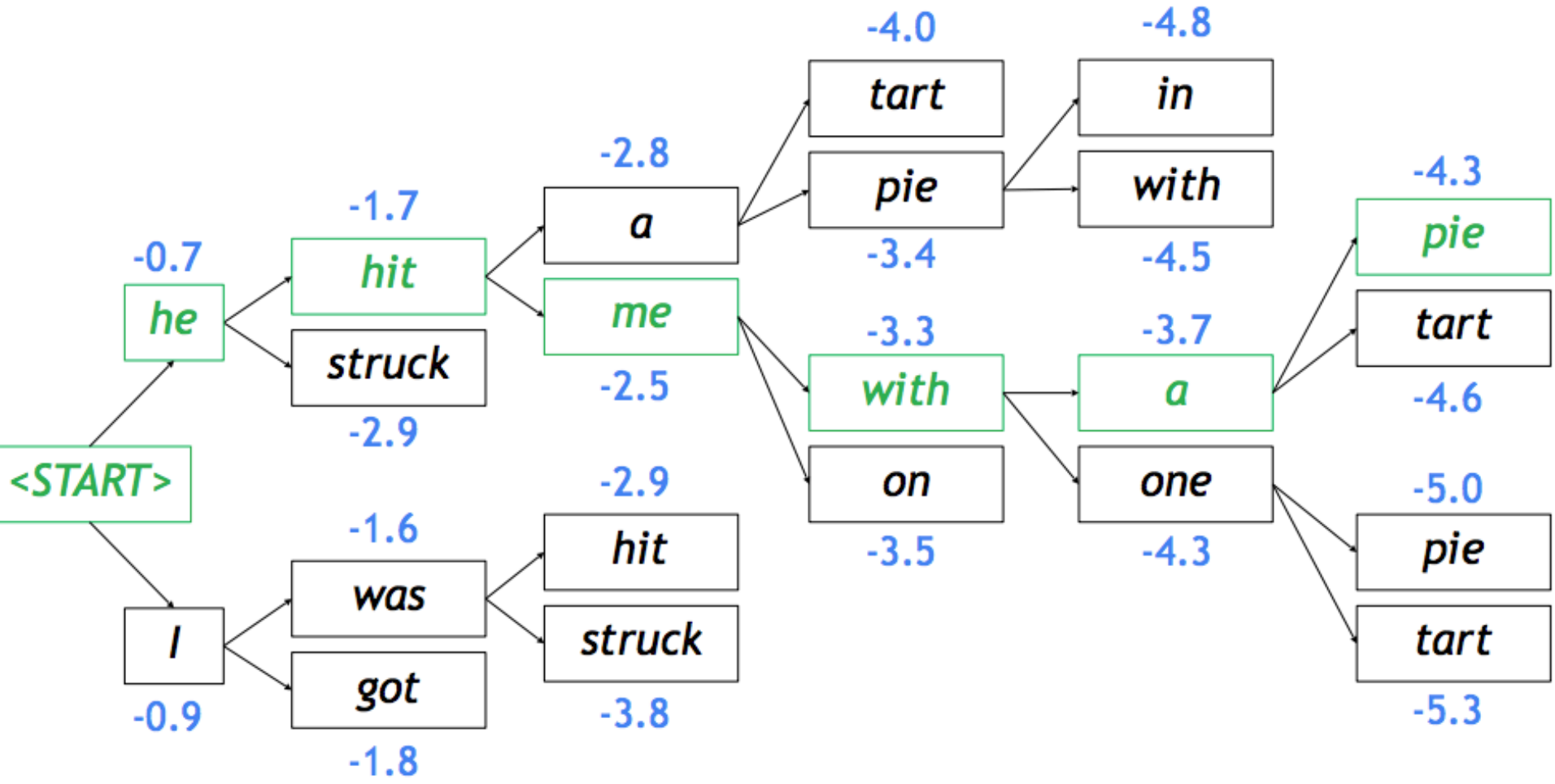
Beam Search: Example



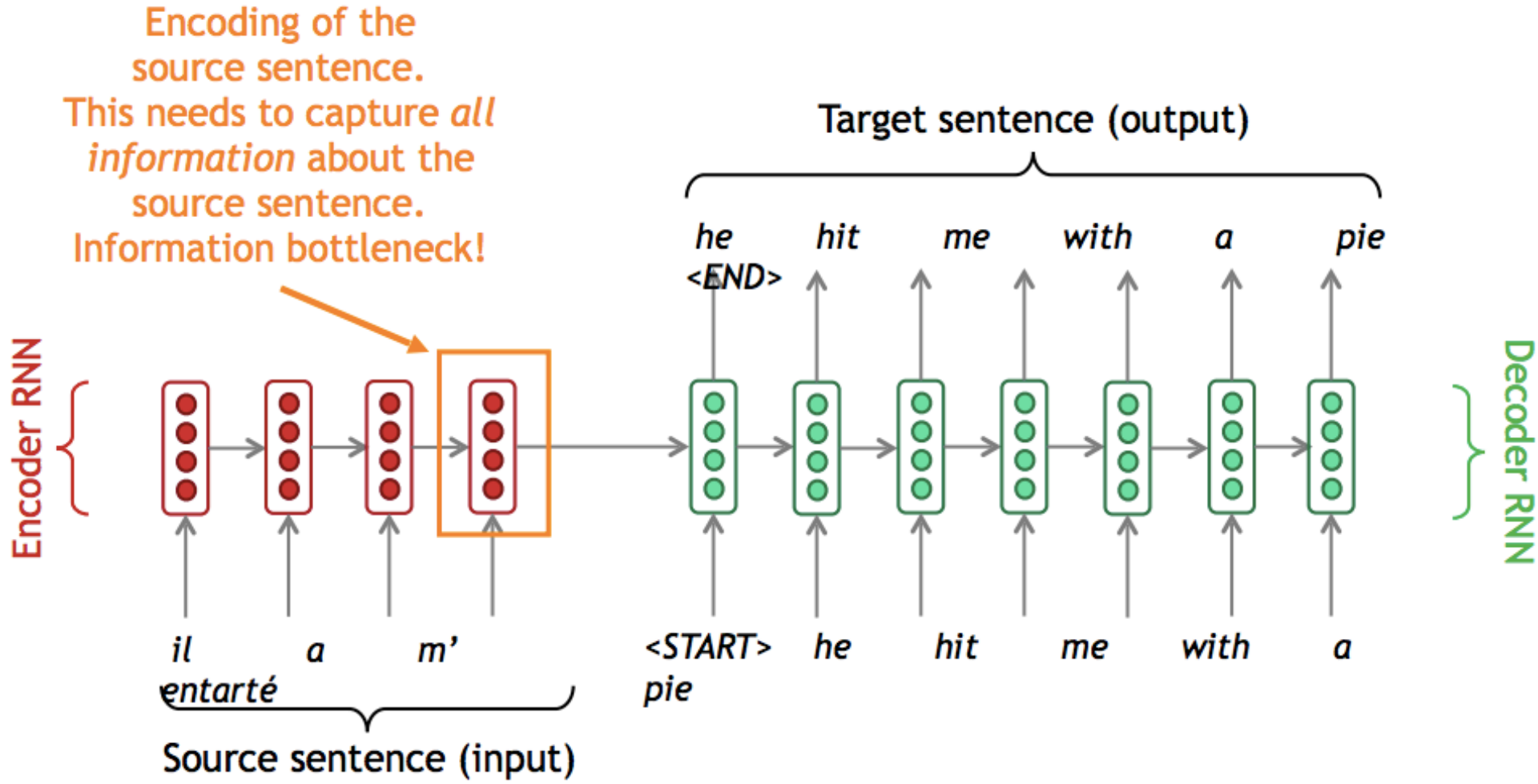
Beam Search: Example



Beam Search: Example



Limitação do Sequence-to-Sequence



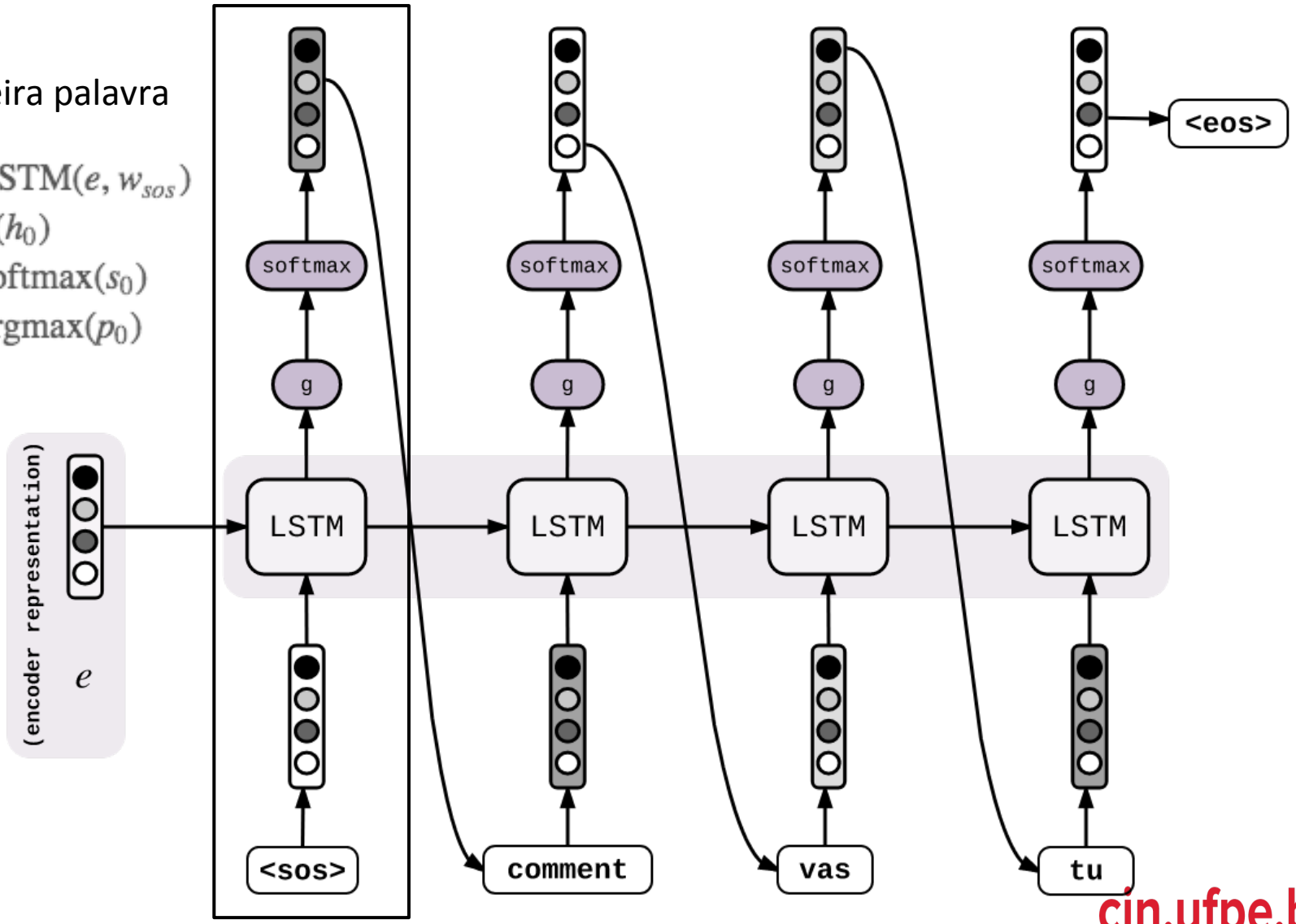
Attention

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

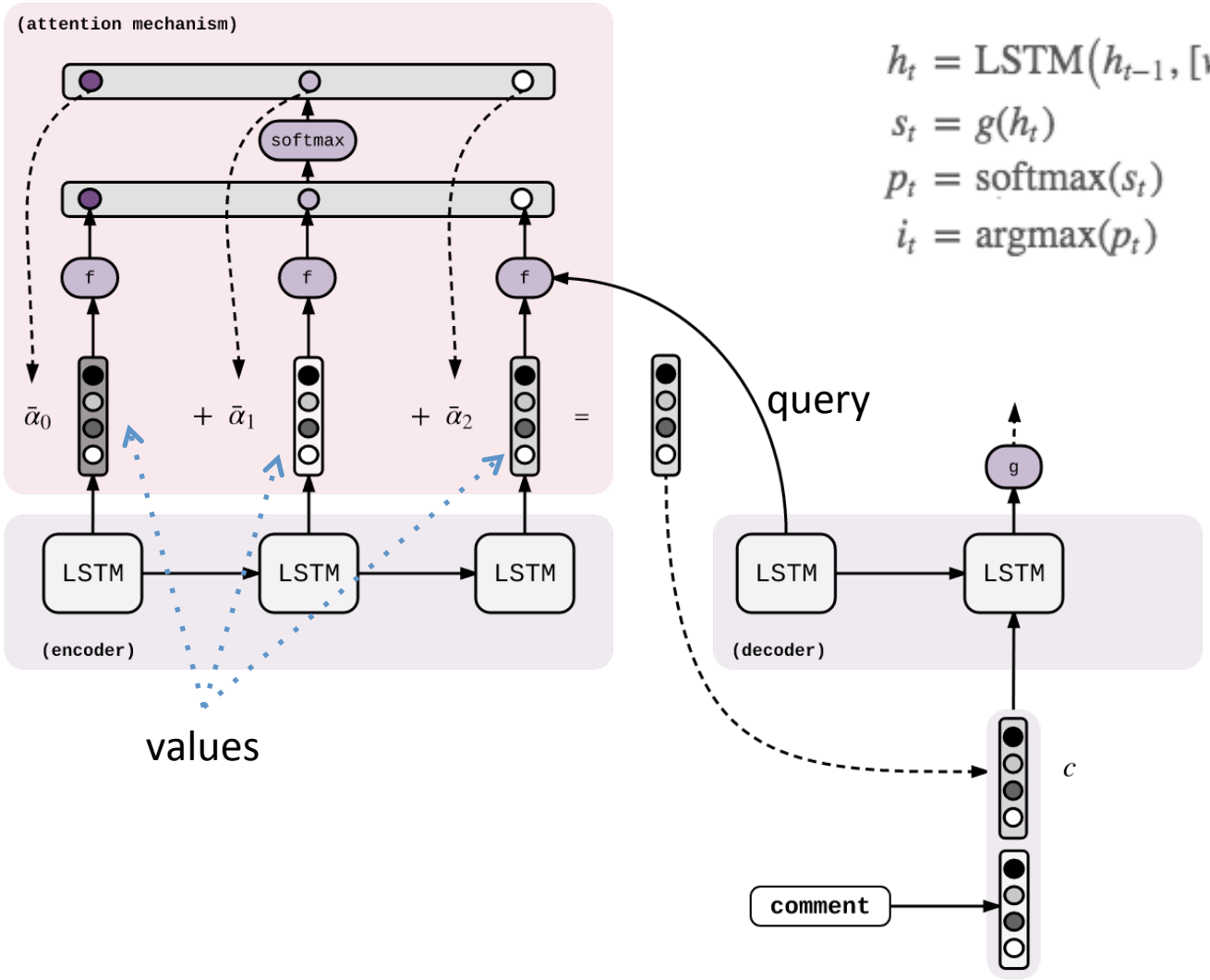
Decoder Tradicional

Primeira palavra

$$h_0 = \text{LSTM}(e, w_{sos})$$
$$s_0 = g(h_0)$$
$$p_0 = \text{softmax}(s_0)$$
$$i_0 = \text{argmax}(p_0)$$



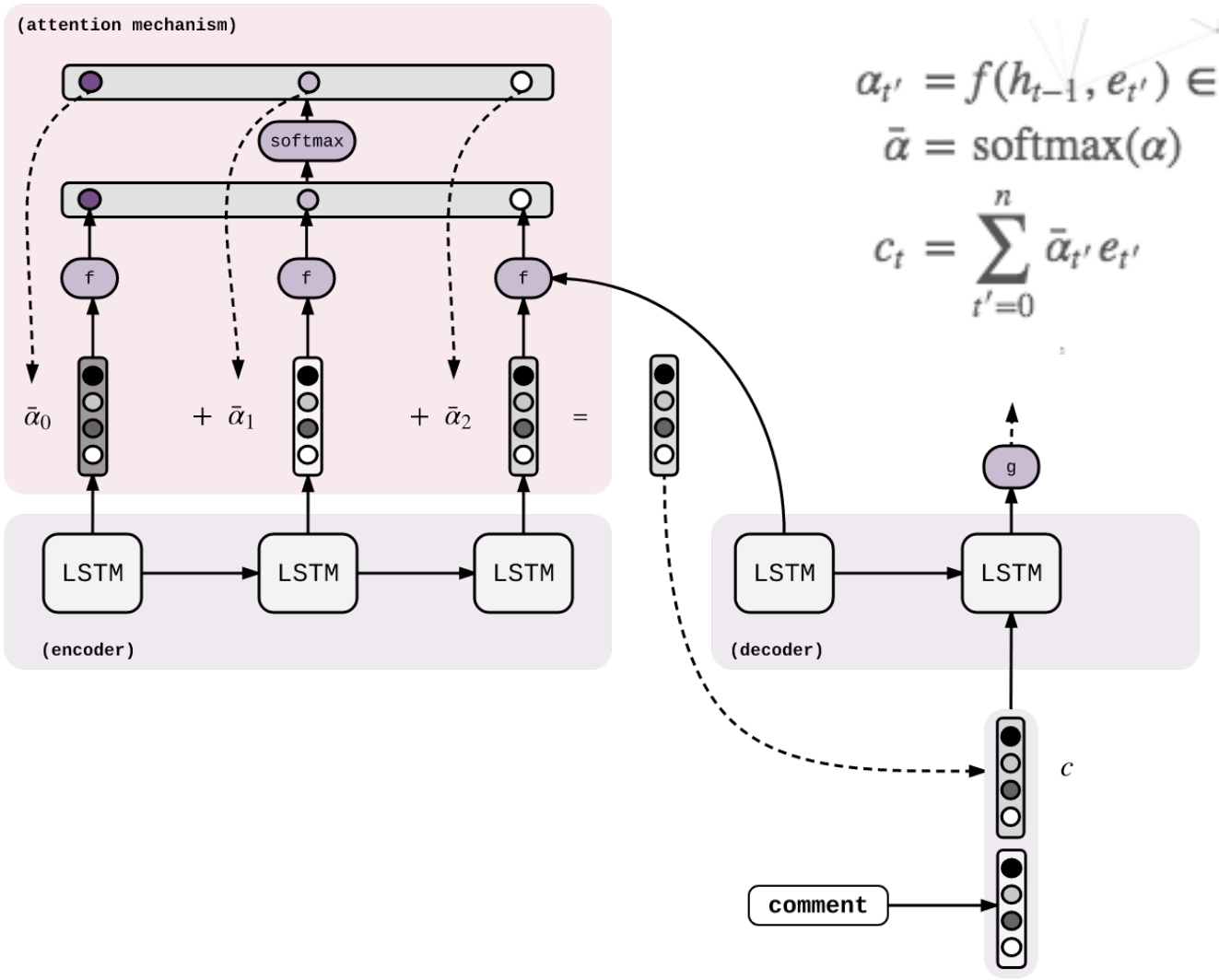
Decoder com Attention



$$h_t = \text{LSTM}\left(h_{t-1}, [w_{i_{t-1}}, c_t]\right)$$
$$s_t = g(h_t)$$
$$p_t = \text{softmax}(s_t)$$
$$i_t = \text{argmax}(p_t)$$

https://tomaxent.com/2018/01/21/Seq2Seq-with-Attention-and-Beam-Search-Repost/

Decoder com Attention



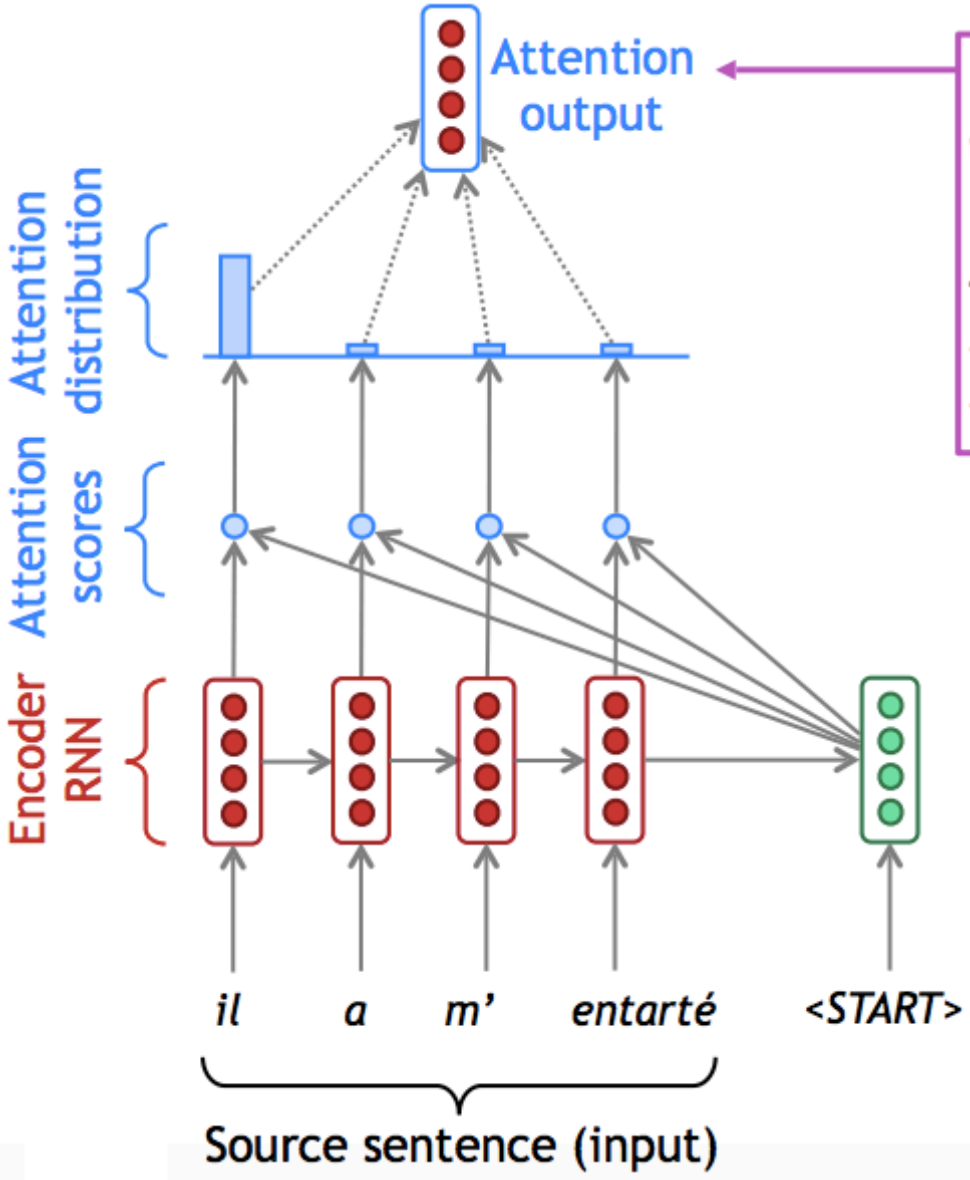
$$\alpha_{t'} = f(h_{t-1}, e_{t'}) \in \mathbb{R} \quad \text{for all } t'$$

$$\bar{\alpha} = \text{softmax}(\alpha)$$

$$c_t = \sum_{t'=0}^{n} \bar{\alpha}_{t'} e_{t'}$$

https://tomaxent.com/2018/01/21/Seq2Seq-with-Attention-and-Beam-Search-Repost/

Sequence-to-Sequence com Attention

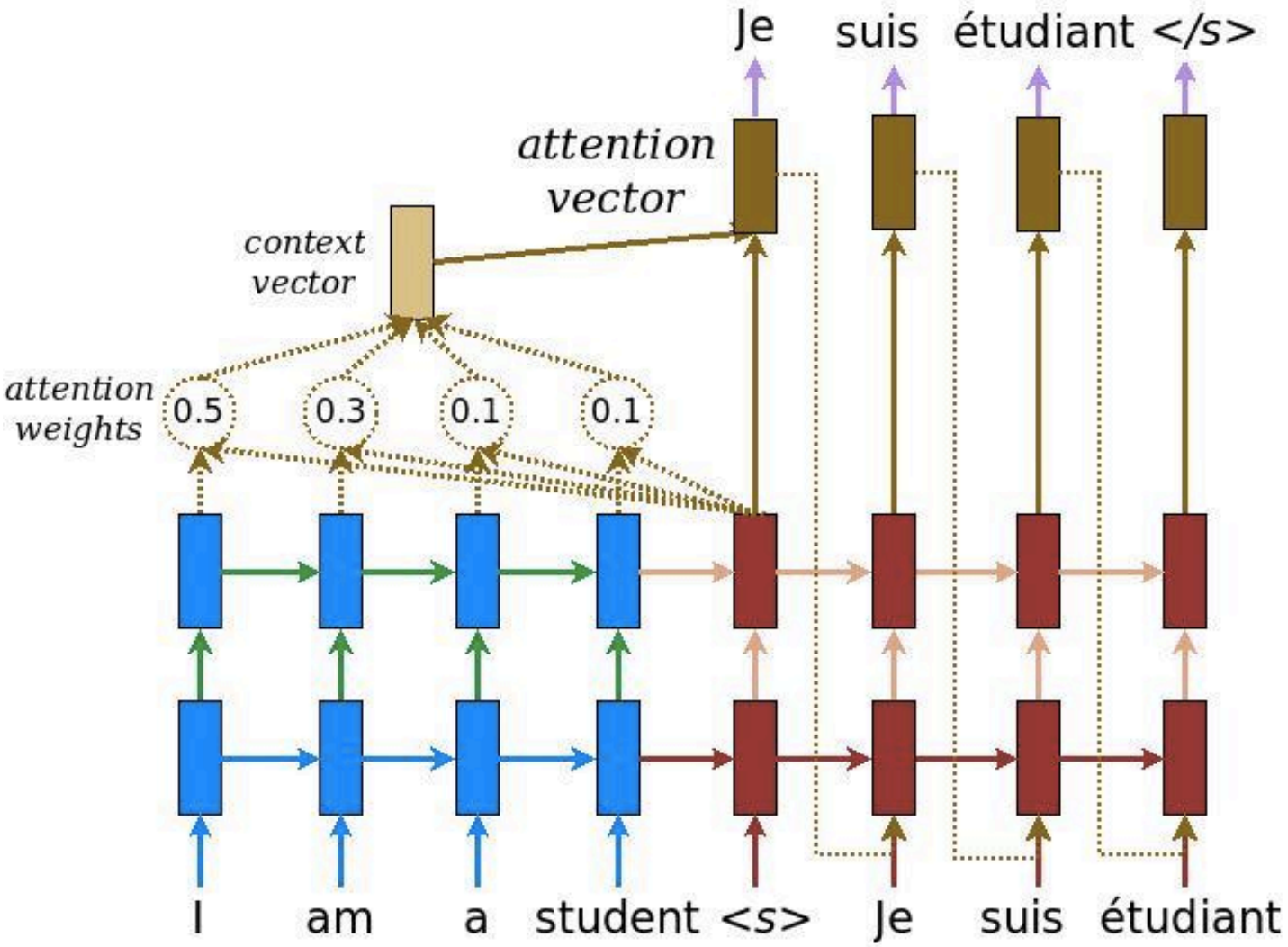


Use the attention distribution to take a **weighted sum** of the encoder hidden states.

The attention output mostly contains information from the hidden states that received high attention.

Decoder RNN

Sequence-to-Sequence com Attention



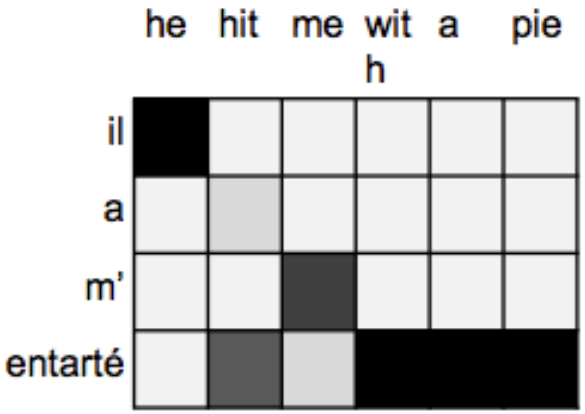
Funções de Attention

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

https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html#summary

Vantages 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



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 (match perfeito)

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

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

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

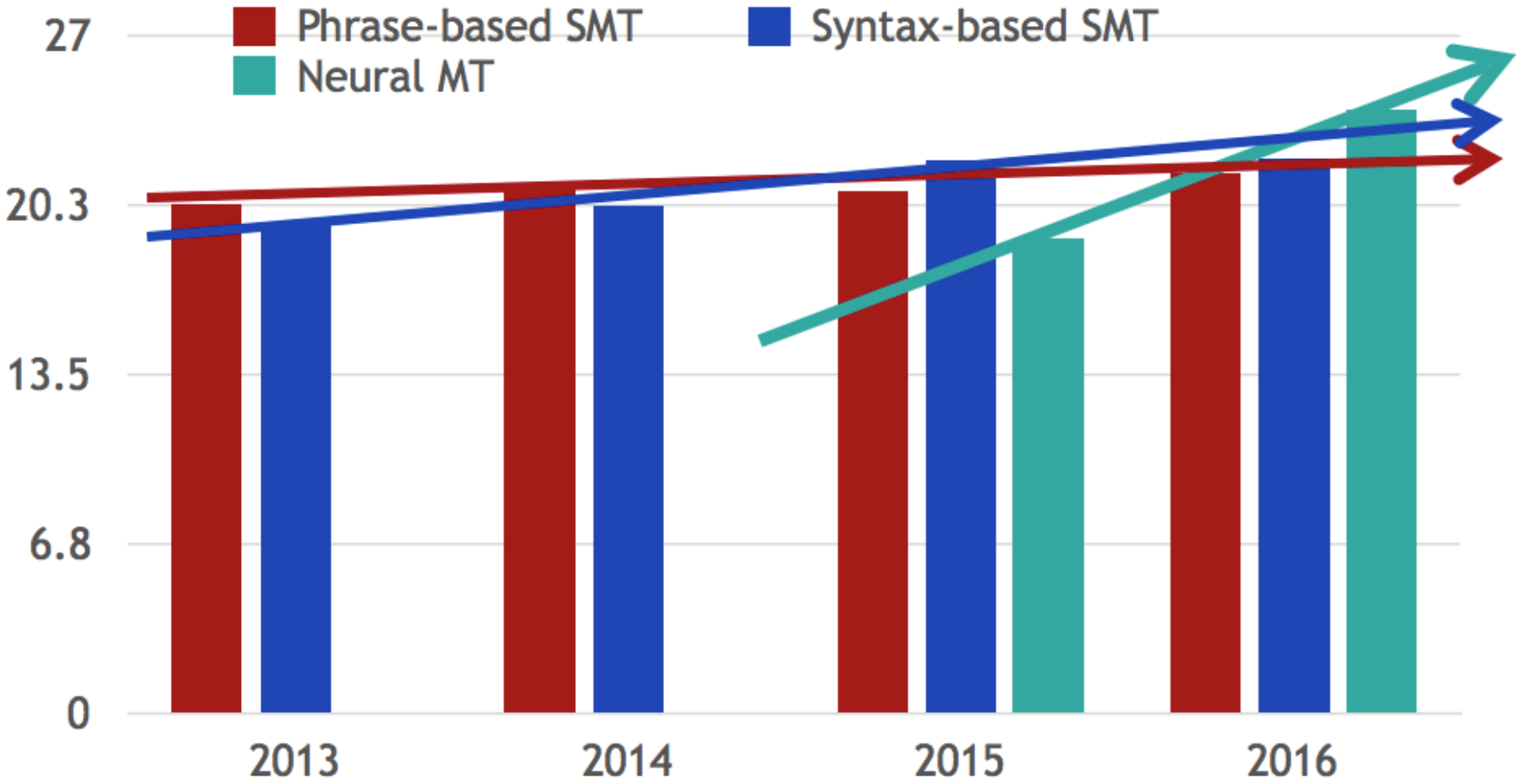
r: referência
c: candidato

https://zhu45.org/posts/2018/Mar/28/bleu-a-method-for-automatic-evaluation-of-machine-translation/

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



Source: http://www.meta-net.eu/events/meta-forum-2016/slides/09_sennrich.pdf

Problemas

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