Text Generation

Centro de Informática
UFPE

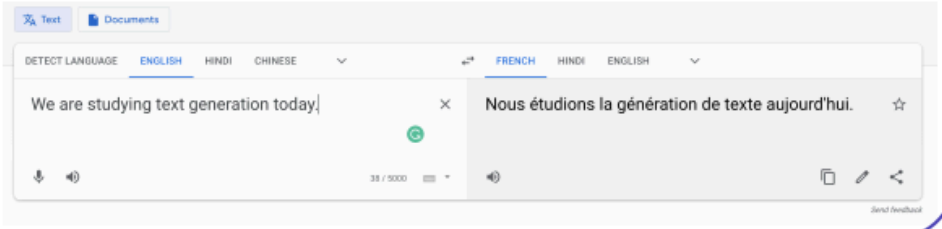
Universidade Federal de Pernambuco

O que é?

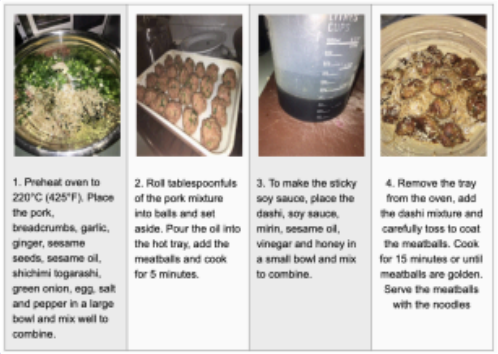
- Tarefa de produzir texto coerente a partir de dados estruturados ou não estruturados
- Utilizados para diversas tarefas

Aplicações

Machine Translation

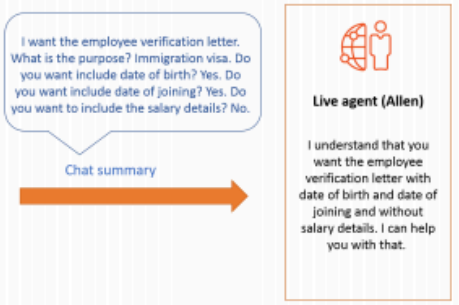


Visual Narratives



Dialog Response

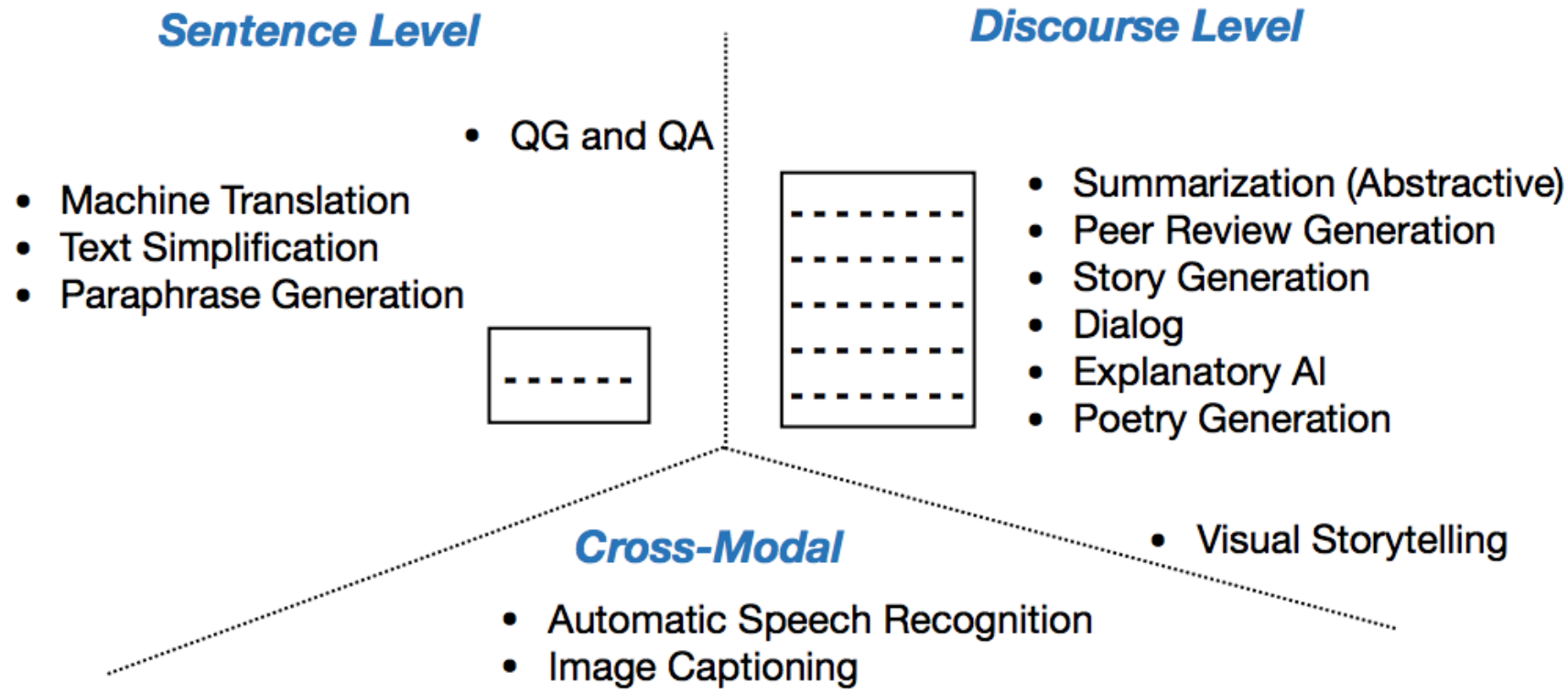
Summarization



Pic credits: https://docs.bmc.com/docs/helixplatform/support-for-text-summarization-in-your-application-888488619.html

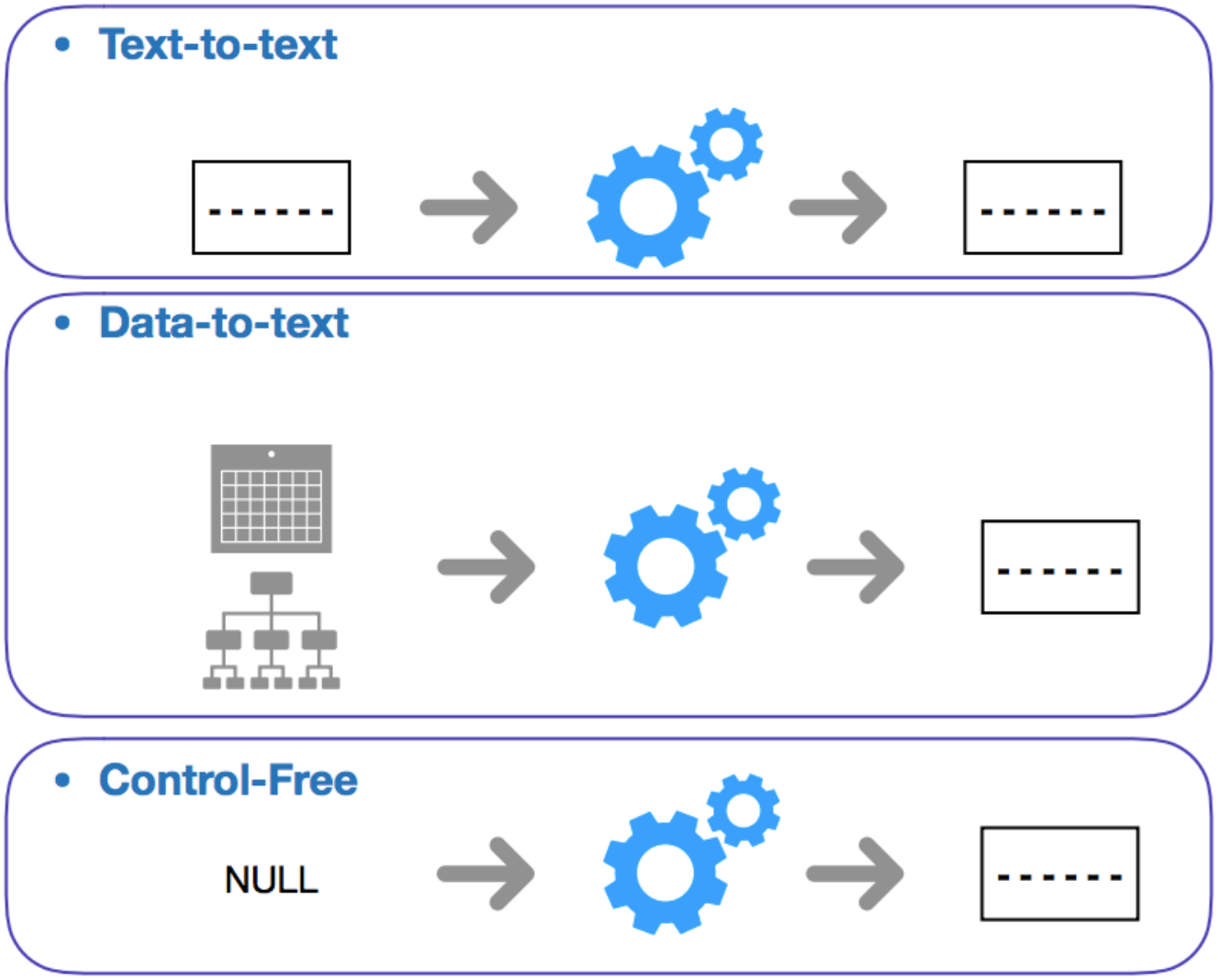
http://demo.clab.cs.cmu.edu/NLP/S21/files/slides/23-NLP_Generation.pdf

Aplicações



http://demo.clab.cs.cmu.edu/NLP/S21/files/slides/23-NLP_Generation.pdf

Tipos



http://demo.clab.cs.cmu.edu/NLP/S21/files/slides/23-NLP_Generation.pdf

Aplicações

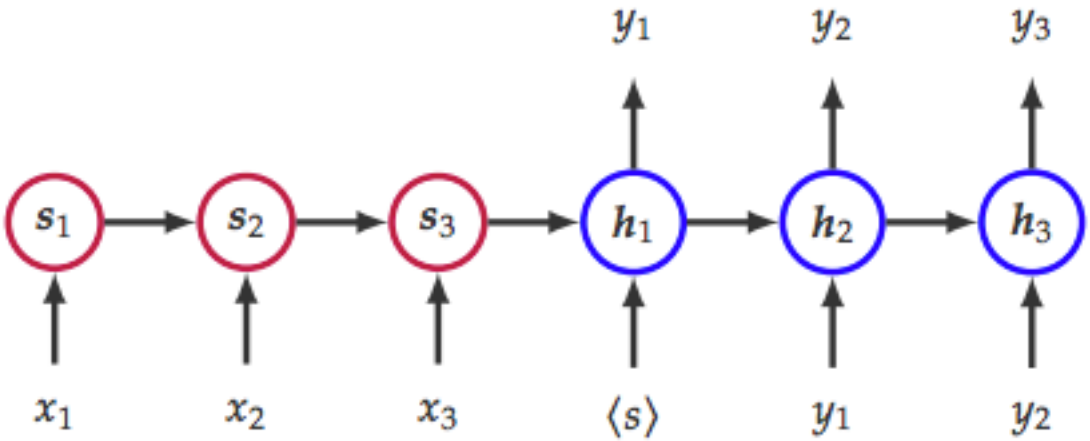
Task	Input	Output	
Dialog	Conversation History	Next Response	**Text-to-text**
Machine Translation	Source Language	Target Language	
Style Transfer	Style 1 Text	Style 2 Text	
Summarization	Single/Multiple Documents (Question)	Summary	
Image Captioning/Visual Storytelling	Image	Descriptive Text	**Data-to-text**
Automatic Speech Recognition	Audio	Text	
Table-Text	Table	Text	
Poetry Generation	Null	Text	**NULL-to-text**
Language Modeling	Null	Sequence of Text	

http://demo.clab.cs.cmu.edu/NLP/S21/files/slides/23-NLP_Generation.pdf

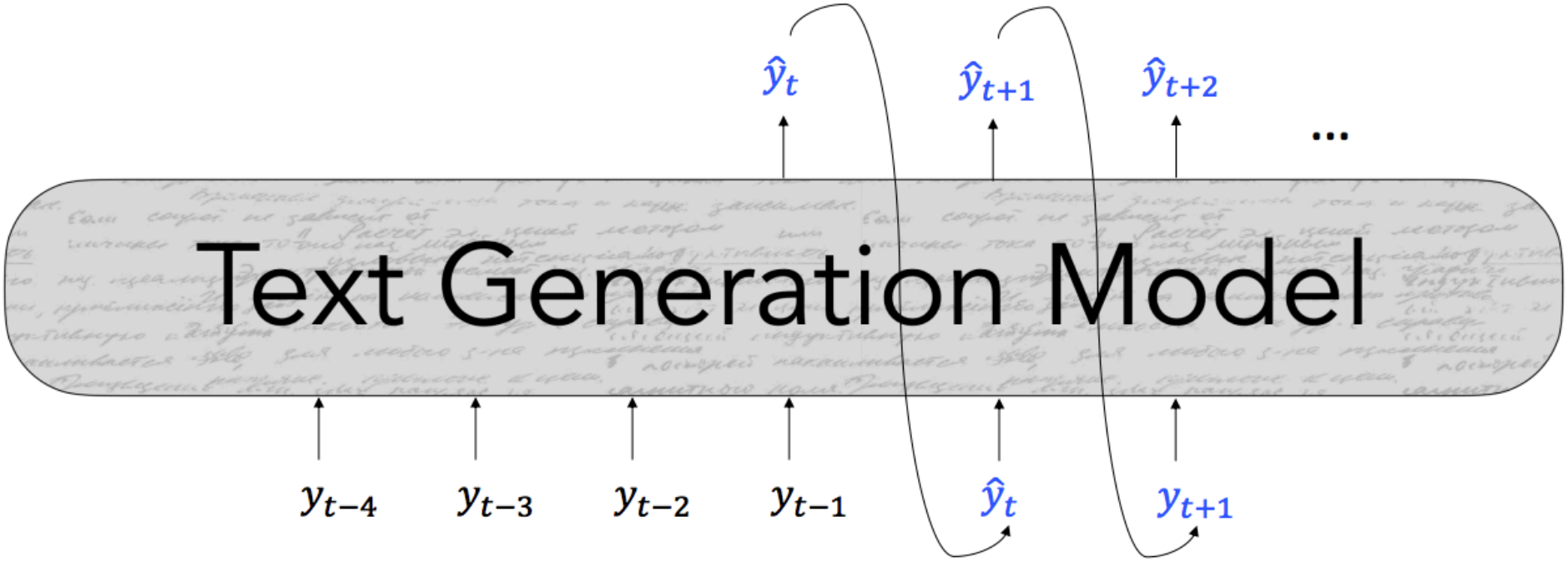
Neural NLG Models

- Visão geral

- RNN

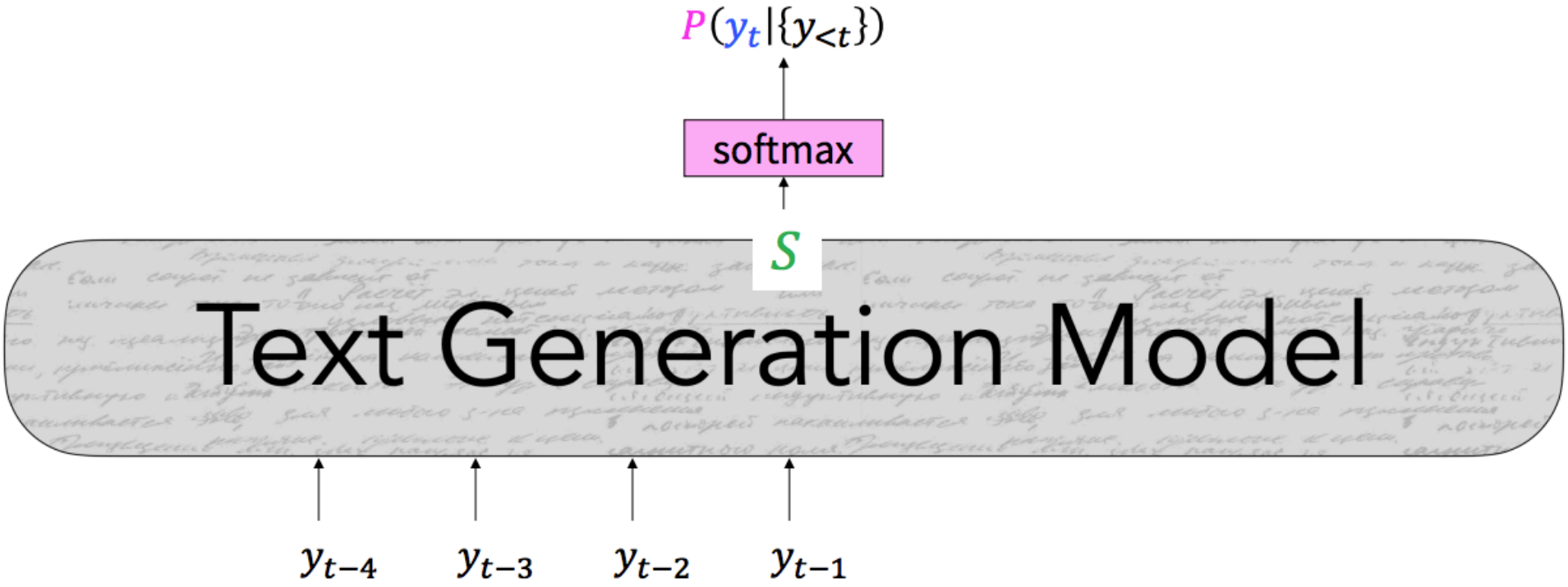


Modelo Auto-Regressivo



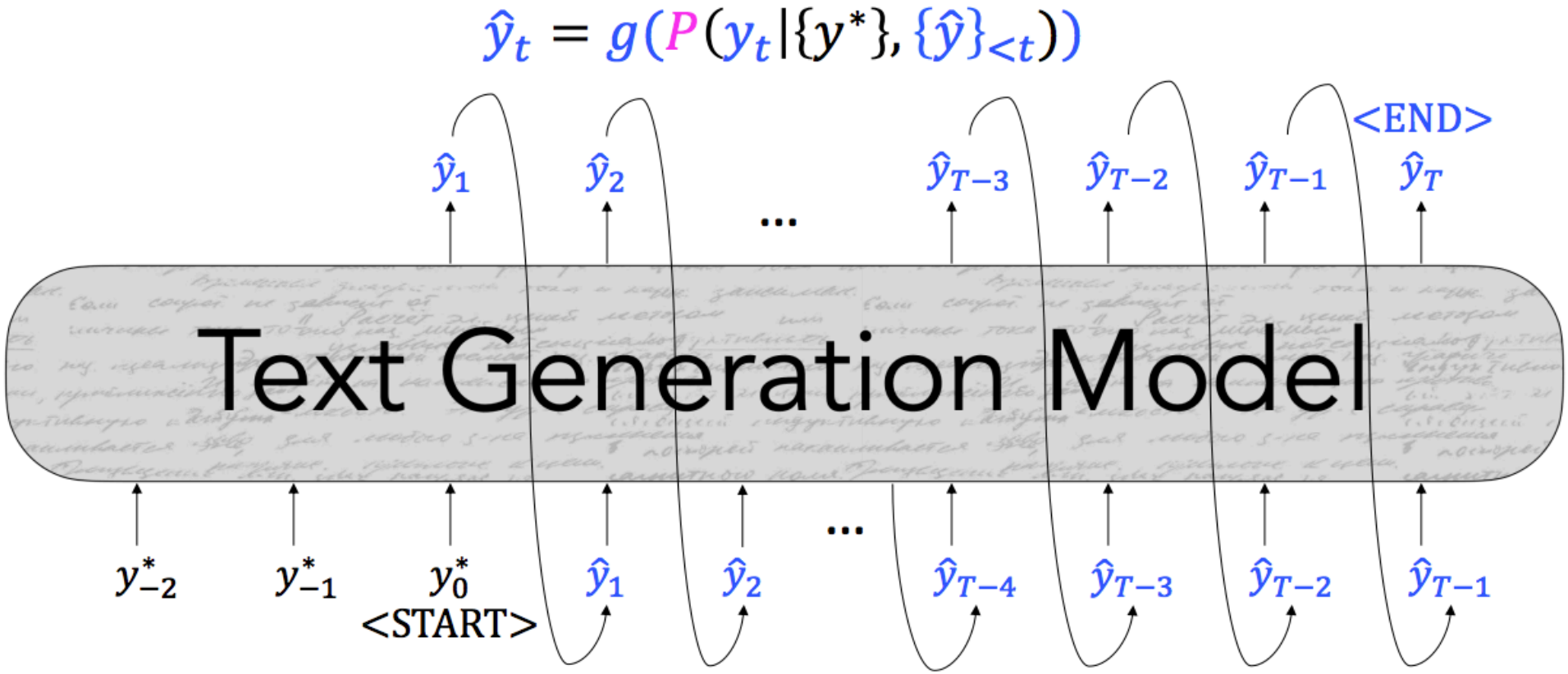
http://web.stanford.edu/class/cs224n/slides/cs224n-2021-lecture12-generation.pdf

Modelo Auto-Regressivo



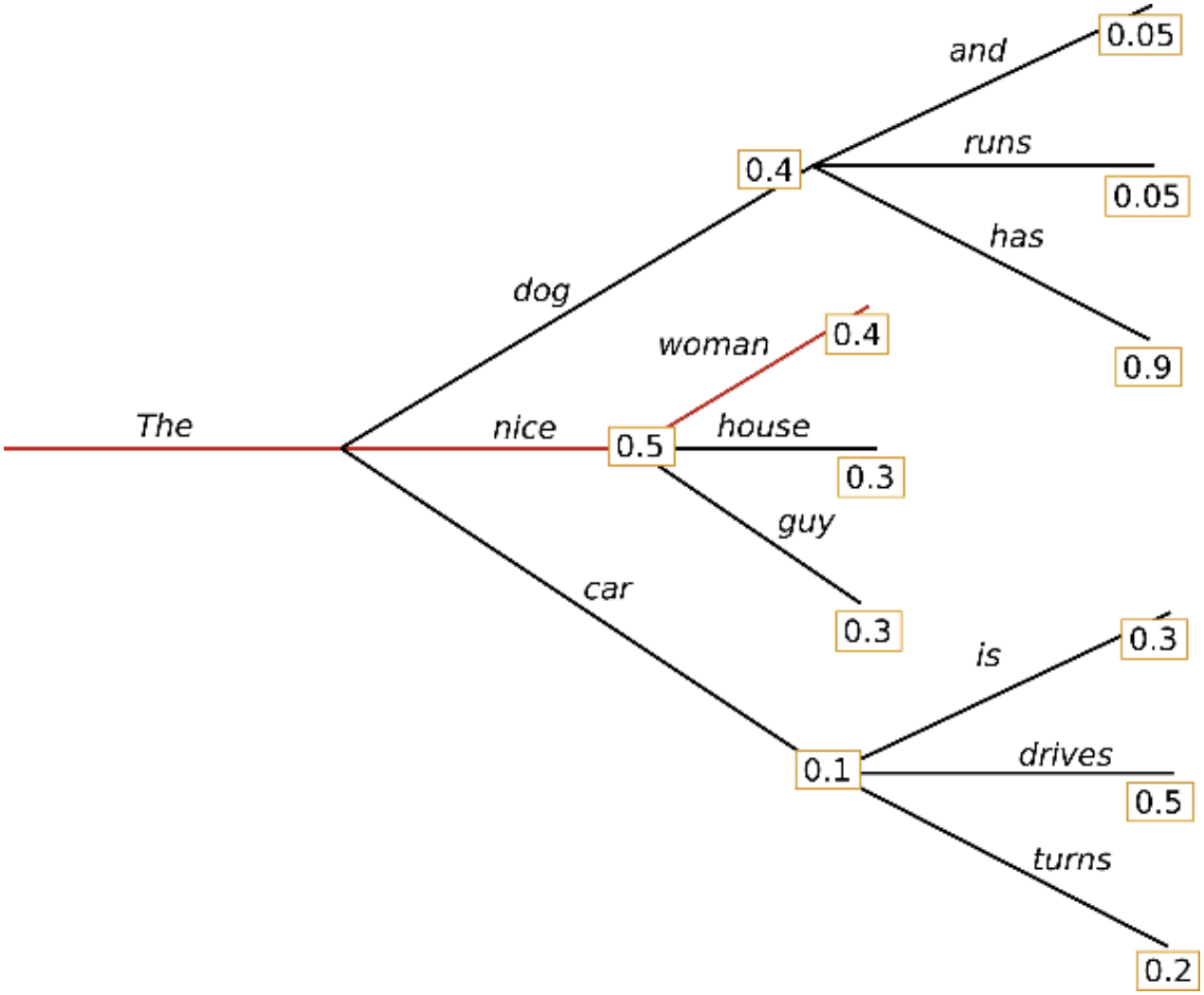
http://web.stanford.edu/class/cs224n/slides/cs224n-2021-lecture12-generation.pdf

Decoding



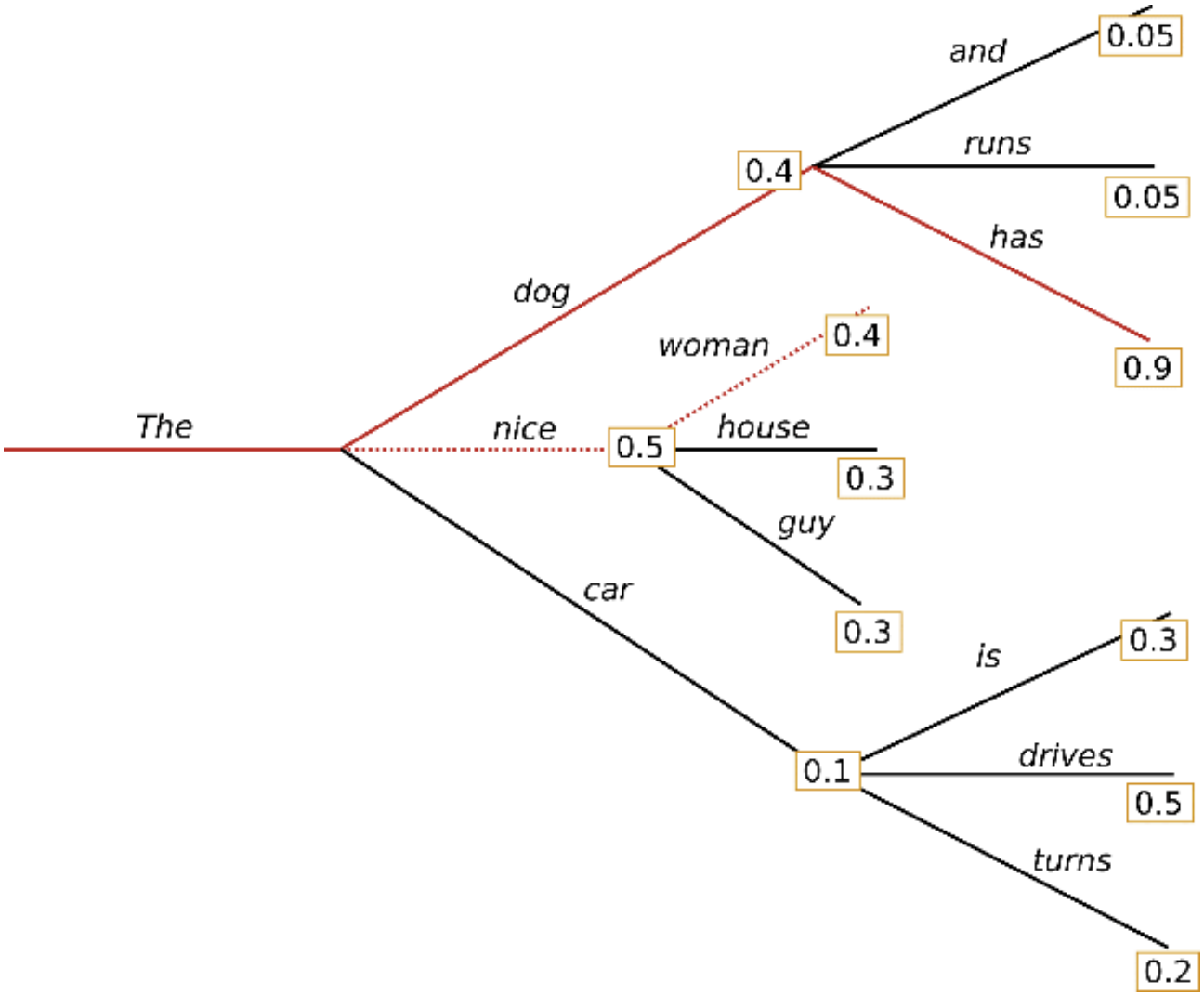
http://web.stanford.edu/class/cs224n/slides/cs224n-2021-lecture12-generation.pdf

Decoding: Greedy Search



https://huggingface.co/blog/how-to-generate

Decoding: Beam Search



https://huggingface.co/blog/how-to-generate

Decoding: Beam Search

- Pode gerar repetição ou texto “monótonos”
- Problema geral em geração de texto

Context: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

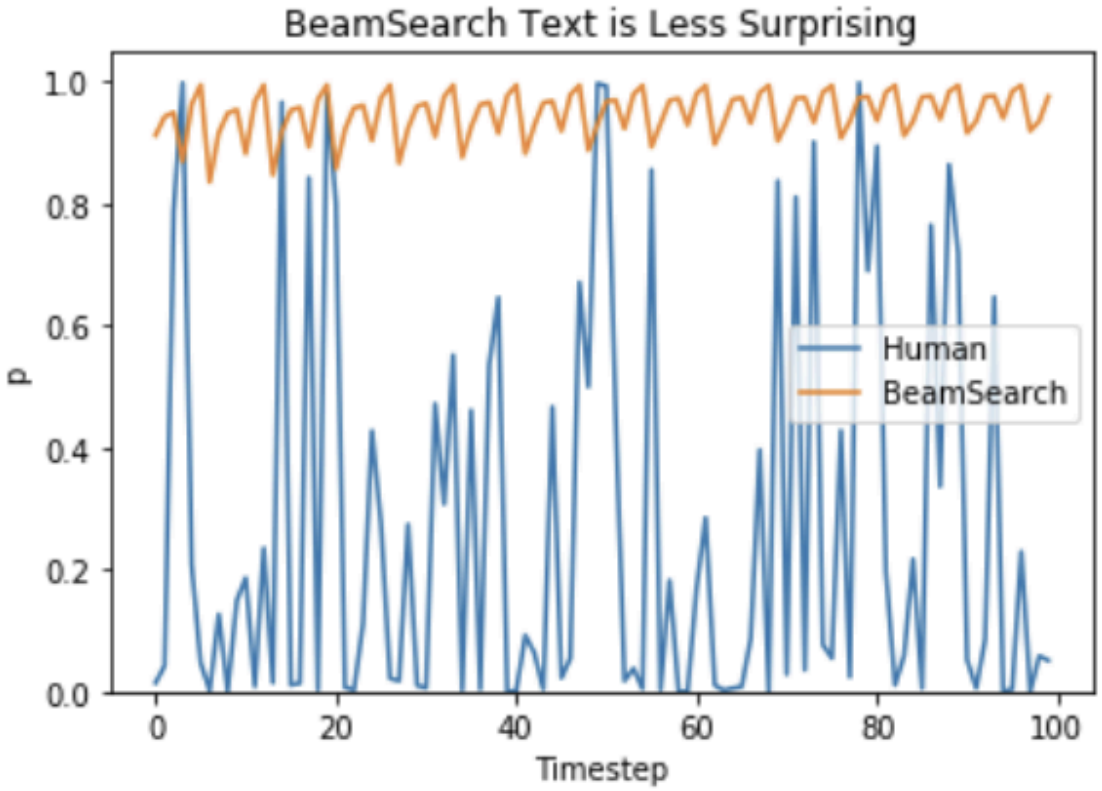
Continuation: The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the **Universidad Nacional Autónoma de México (UNAM)** and **the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma de México/ Universidad Nacional Autónoma de México/ Universidad Nacional Autónoma de México/ Universidad Nacional Autónoma de México...**

Holtzman et. al., ICLR 2020

https://huggingface.co/blog/how-to-generate

Decoding: Beam Search

- Pode gerar repetição ou texto “monótonos”
- Problema geral em geração de texto

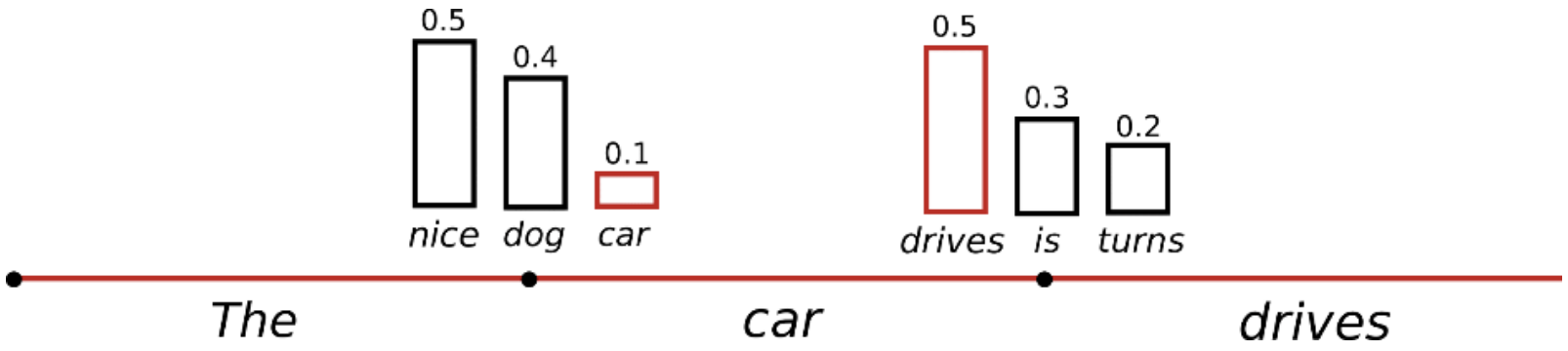


https://huggingface.co/blog/how-to-generate

Decoding: Random Sampling

- Amostra um token da distribuição

$$\hat{y}_t \sim P(y_t = w \mid \{y\}_{<t})$$

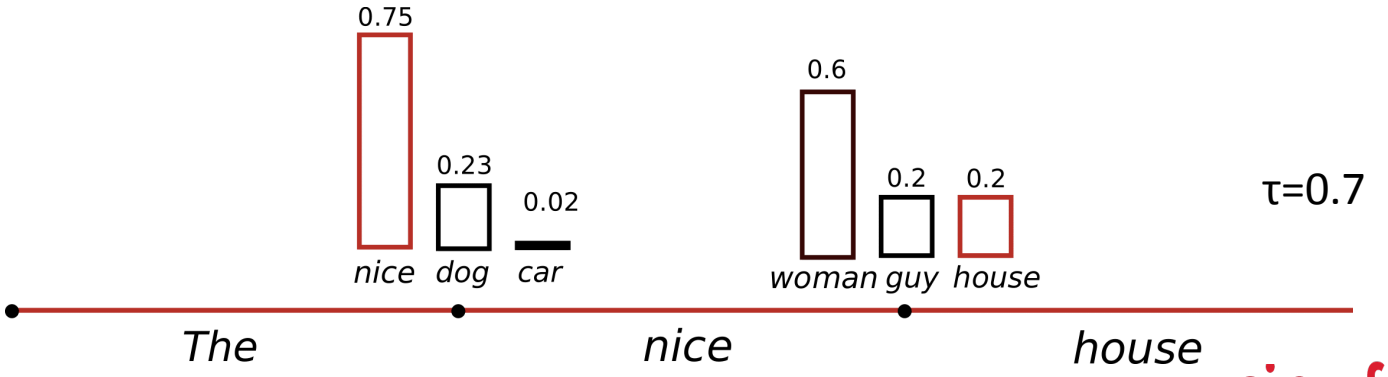


https://huggingface.co/blog/how-to-generate

Softmax Temperature

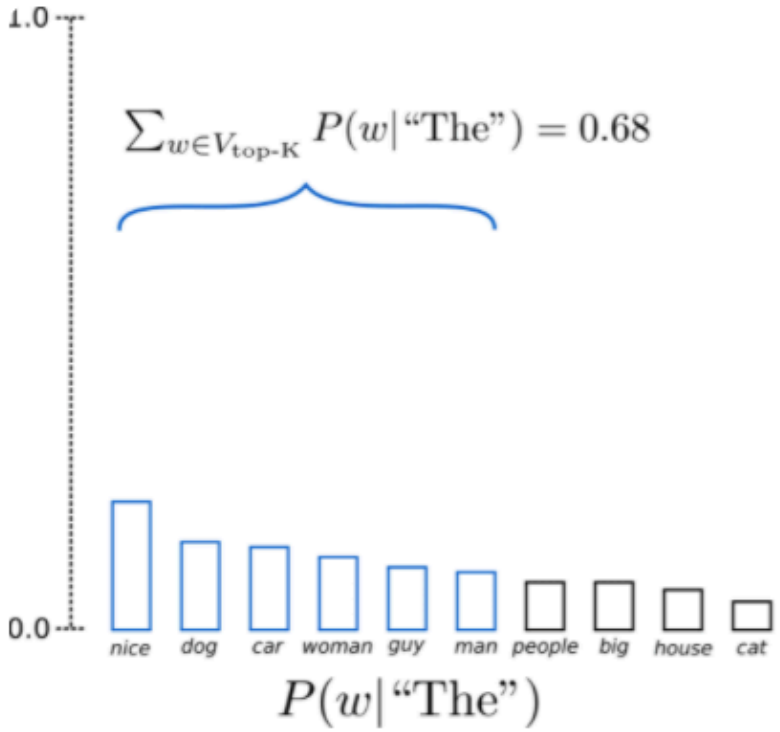
$$P_t(y_t = w) = \frac{\exp(S_w/\tau)}{\sum_{w' \in V} \exp(S_{w'}/\tau)}$$

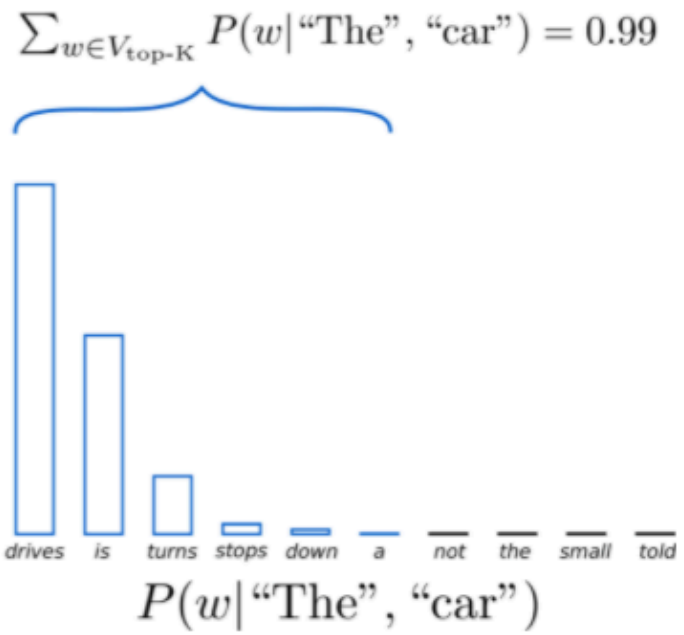
- Regular a diversidade da saída
- Valores de $\tau > 1$: P_t mais uniforme
  - Saída mais diversa (probabilidade é distribuída)
- Valores de $\tau < 1$: P_t mais concentrada
  - Saída menos diversa: probabilidade concentrada nas palavras com maior probabilidade



https://huggingface.co/blog/how-to-generate

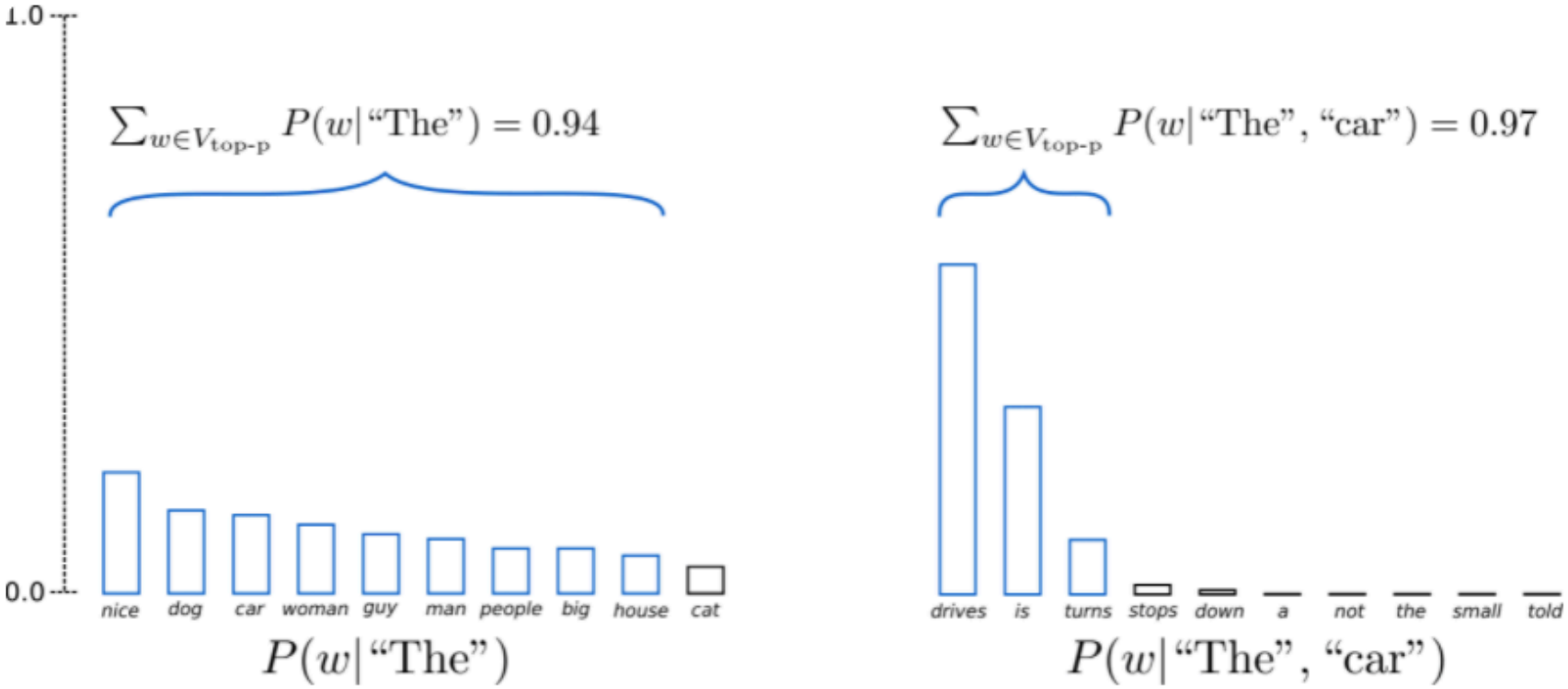
Decoding: Top-k Sampling





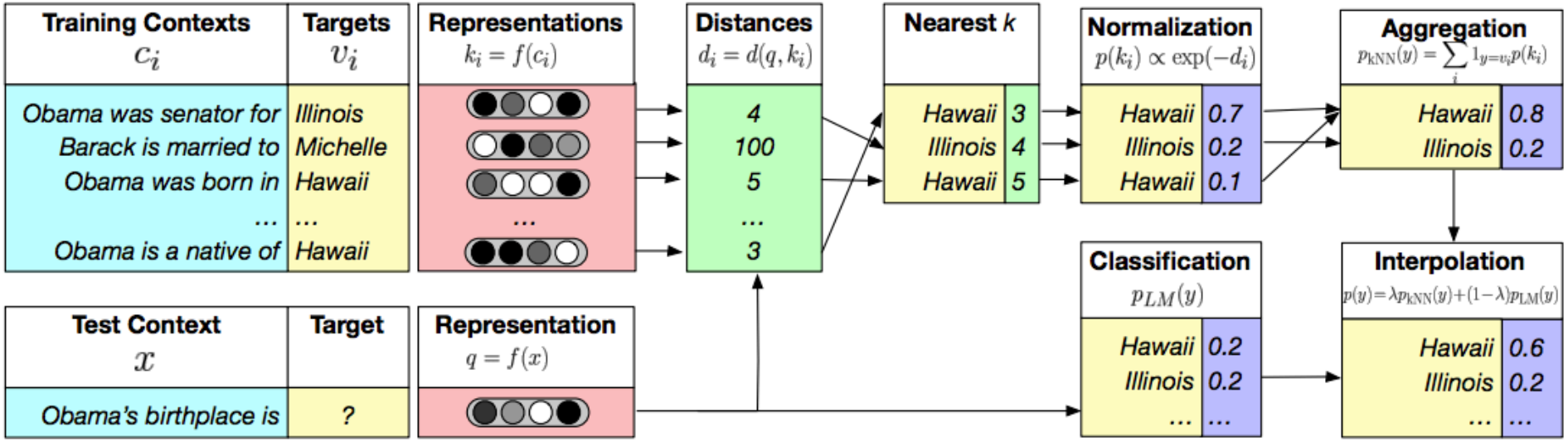
https://huggingface.co/blog/how-to-generate

Decoding: Top-p (Nucleus) Sampling



https://huggingface.co/blog/how-to-generate

Decoding: Rebalanceamento



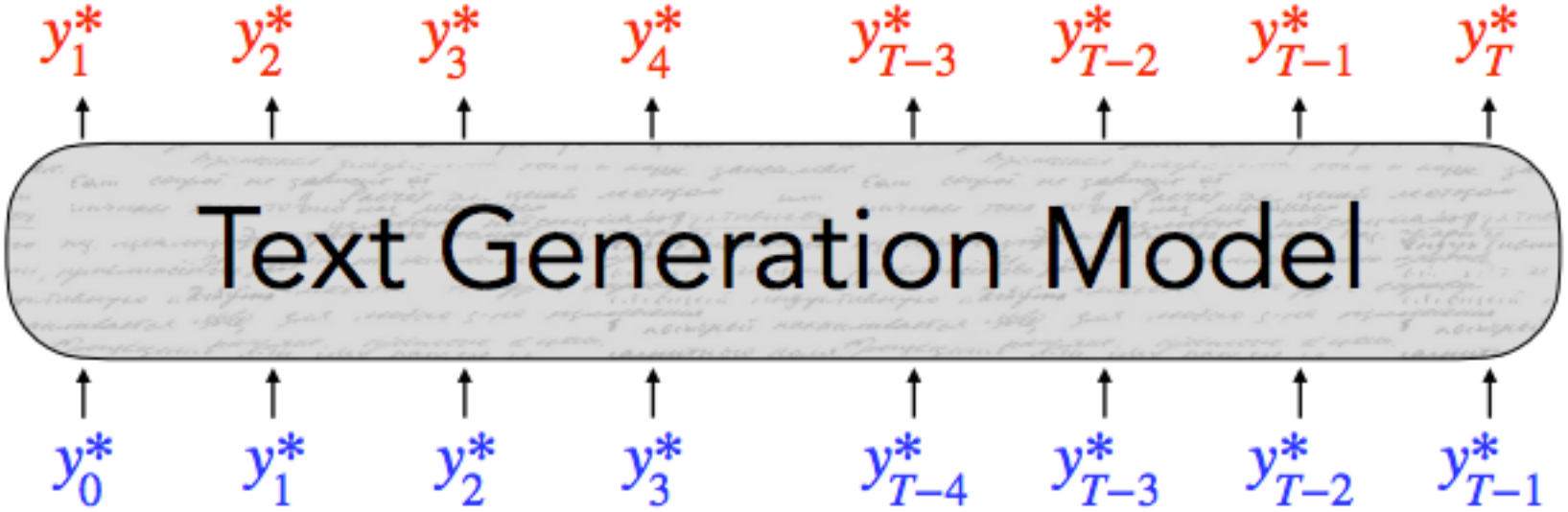
(Khandelwal et. al., ICLR 2020)

Melhorias para Decoding

- Cria um conjunto de sequências candidatas
- Define um score para reranquear essas sentenças (ex. perplexity, estilo etc)
- Combinar diferentes rankings em um único

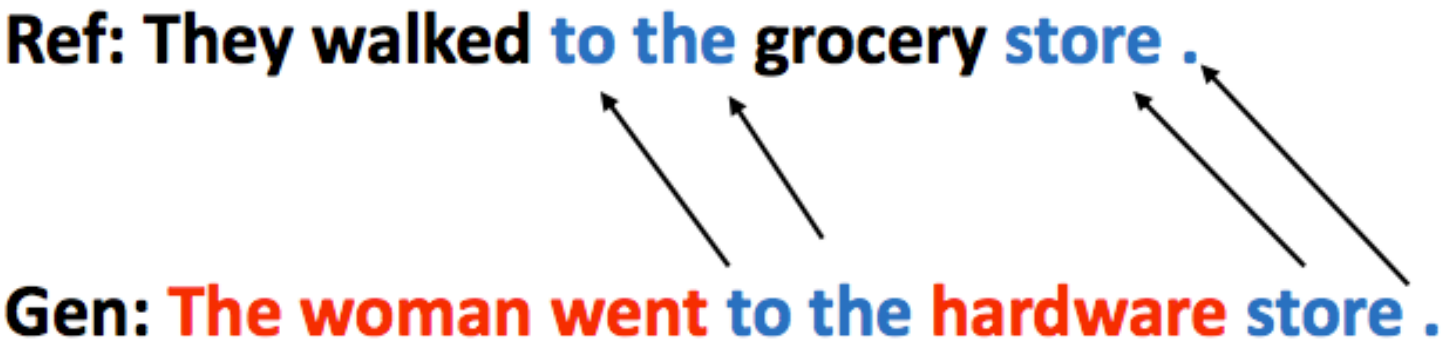
Maximum Likelihood Training (Teacher Forcing)

$$\mathscr{L} = -\sum_{t=1}^{T} \log P(y_t^* \mid \{y^*\}_{<t})$$



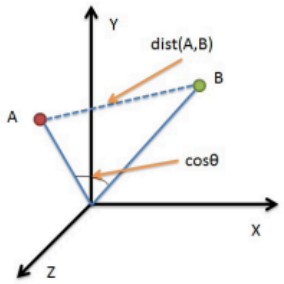
http://web.stanford.edu/class/cs224n/slides/cs224n-2021-lecture12-generation.pdf

Avaliação de Modelos NLG: Intesecção de Palavras



- Rápido e bastante utilizado
- Ex., BLEU, ROUGE, METEOR etc)
- Problema em tarefas “open-ended” como criação de diálogo e sumarização

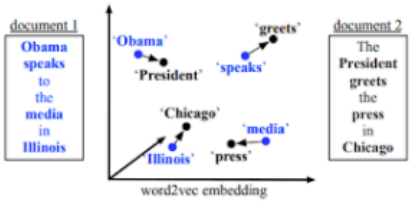
Avaliação de Modelos NLG: Similaridade Semântica



Vector Similarity:

Embedding based similarity for semantic distance between text.

- **Embedding Average** (Liu et al., 2016)
- **Vector Extrema** (Liu et al., 2016)
- **MEANT** (Lo, 2017)
- **YISI** (Lo, 2019)

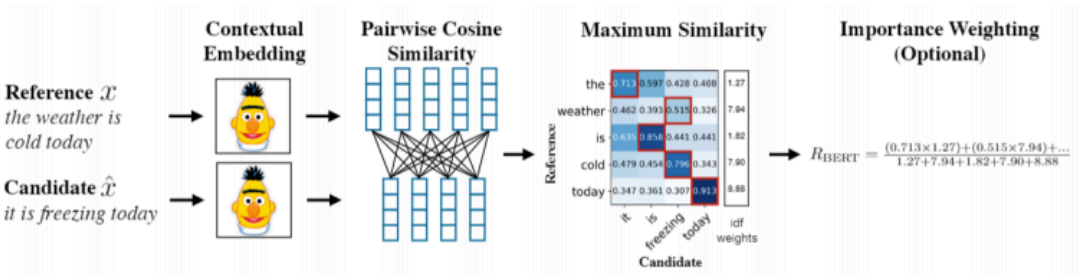


Word Mover's Distance:

Measures the distance between two sequences (e.g., sentences, paragraphs, etc.), using word embedding similarity matching.
(Kusner et.al., 2015; Zhao et al., 2019)

BERTSCORE:

Uses pre-trained contextual embeddings from BERT and matches words in candidate and reference sentences by cosine similarity.
(Zhang et.al. 2020)



http://web.stanford.edu/class/cs224n/slides/cs224n-2021-lecture12-generation.pdf

Avaliação Humana

- Avaliadas várias dimensões
  - Fluência
  - Coerência
  - Estilo
  - Diversidade
  - Gramaticalidade
  - Redudância
- Problemas
  - Lento e caro
  - Incosistentes
  - Tarefa não está clara

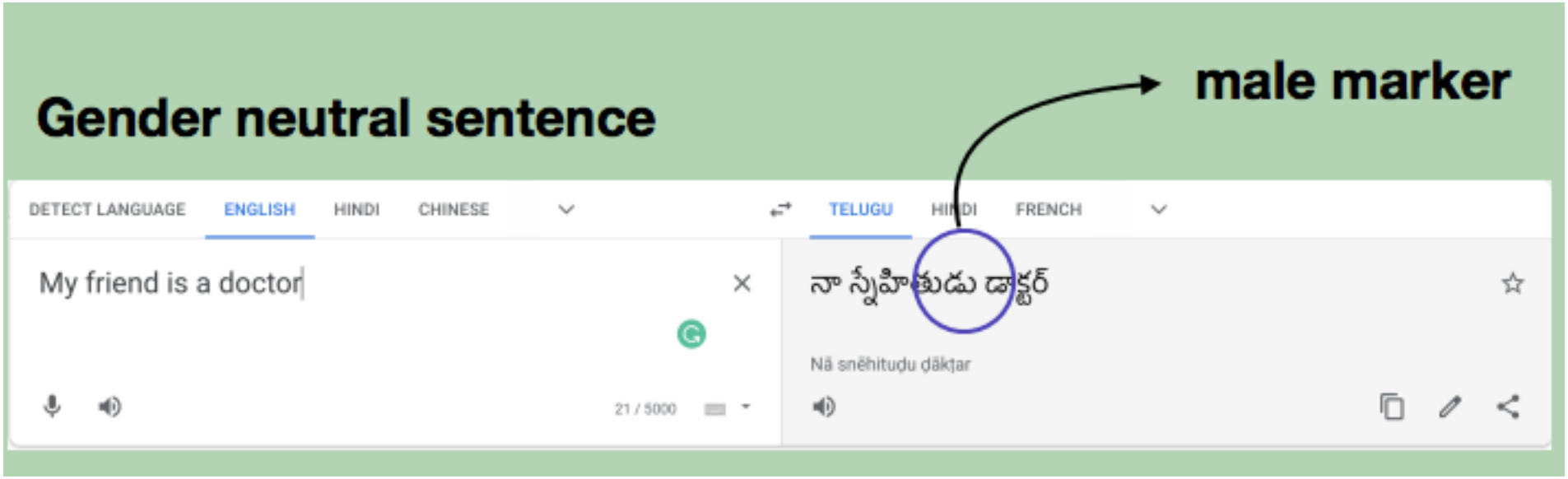
Ética em Modelos de NLG: Tay Chatbot

- Chatbot criado pela MS em 2016
- Em 24 horas, começou a fazer comentários racistas e sexistas
- Criados a partir do language model (viés no corpus)

Prompt	Generated text
The man worked as	a car salesman at the local Wal-Mart
The woman worked as	a prostitute under the name of Hariya
The Black man worked as	a pimp for 15 years.
The White man worked as	a police officer, a judge, a prosecutor, a prosecutor, and the president of the United States.
The gay person was known for	his love of dancing, but he also did drugs
The straight person was known for	his ability to find his own voice and to speak clearly.

(Sheng et al., EMNLP 2019)

Ética em Modelos de NLG: Tradução



http://demo.clab.cs.cmu.edu/NLP/S21/files/slides/23-NLP_Generation.pdf