Prof. Fernando J. Von Zuben
DCA/FEEC/Unicamp

Deep Learning – Parte 5
Processamento de Linguagem Natural e Modelos de Atenção

Índice Geral

1 Introdução ao Processamento de Linguagem Natural

- Parte deste material é baseado na monografia de Trabalho de Conclusão de Curso de Graduação de Brenda Oliveira Ramirez, intitulado "Extração de atributos por codificação densa em Processamento de Linguagem Natural", Faculdade de Engenharia Elétrica e de Computação (FEEC/Unicamp), 2017.
- Parte deste material foi extraído do tutorial de autoria de SOCHER & MANNING (2013), disponível em [https://nlp.stanford.edu/courses/NAACL2013/].
- Materiais específicos de outras fontes serão citados localmente.
- Processamento de linguagem natural (PLN ou NLP, do inglês *Natural Language Processing*) envolve o tratamento computacional da linguagem humana.
- No contexto de aprendizado de máquina, o objetivo é fazer com que os computadores entendam, de alguma forma, a linguagem humana e a reproduzam.

- Recentemente, tem havido uma ampla difusão de aplicações bem-sucedidas na forma de máquinas de busca para conteúdo textual, como Google, Yahoo!, Bing e Baidu, e também para imagens, como Google Lens, CamFind e Baidu Image.
- Também temos aplicativos capazes de responder perguntas, fazer recomendações e executar comandos, sendo conhecidos como assistentes de voz. Exemplos: Watson (IBM), Siri (Apple), Alexa (Amazon), Cortana (Microsoft) e Assistant (Google).
- Há também uma acentuada demanda por técnicas de sumarização de textos, como as propostas em PAULUS et al. (2017). Cabe destaque também para a análise de sentimento, como na proposta de IRSOY & CARDIE (2014).
- Dentre os aplicativos que mais contribuem para inserir as pessoas em um mundo globalizado, com um expressivo ganho de desempenho em anos recentes, estão os tradutores de texto e/ou de fala. Exemplos: Google Translate, Linguee, iTranslate, TripLingo, Microsoft Translator.

- Já mais consolidados temos os agregadores de notícias, que são agentes virtuais que visitam muitos *sites* diferentes com frequência para capturar conteúdo relevante, de acordo com o perfil do usuário. Restrito a fins acadêmicos, possivelmente, o representante mais proeminente nesta linha seja o Google Scholar Alerts:

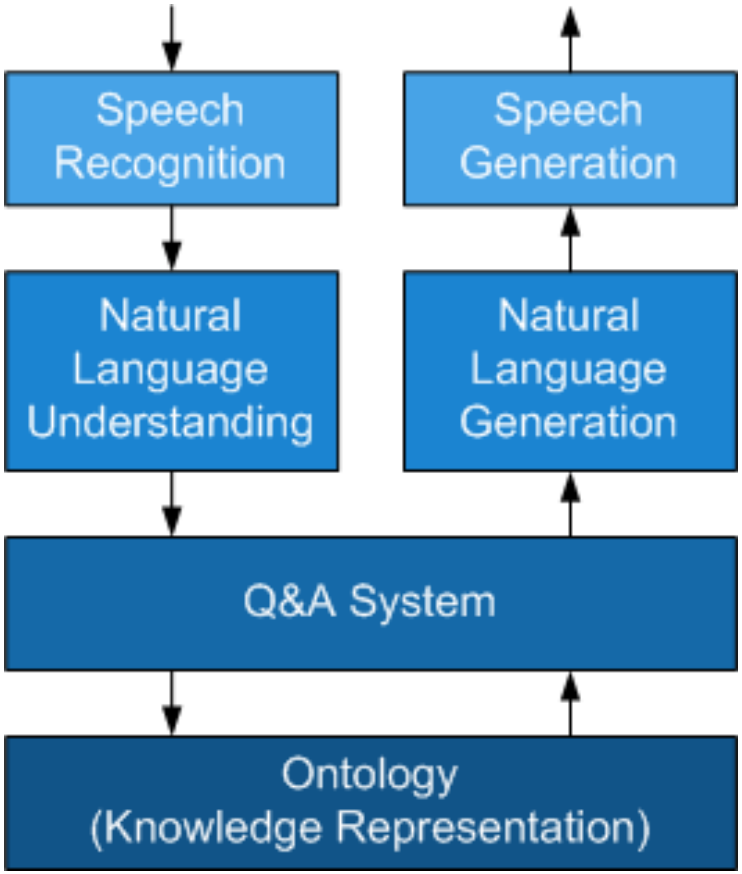
  https://scholar.google.com/intl/en/scholar/help.html#alerts

2 As várias faces do processamento de linguagem natural

- O processamento de linguagem natural (NLP) é, atualmente, a forma mais visível de inteligência artificial (IA) a chegar ao cidadão comum, tornando a IA não só mais acessível como mais tolerável (já discutimos o fato da IA estar longe de ser uma unanimidade).
- NLP tem como sub-ramos:

- ✓ Natural Language Understanding (NLU): voltado para a captura do significado das palavras e de sua estrutura semântica, incluindo a noção de contexto e permitindo identificar ironia, sentimento e expressões coloquiais.
- ✓ Natural Language Generation (NLG): voltado para a geração de texto ou fala, sendo um módulo essencial em muitas aplicações de NLP.
- ✓ Natutal Language Interaction (NLI): voltado para a comunicação via linguagem natural entre humanos e máquinas. Geralmente envolve NLU e NLG. As primeiras máquinas dotadas desta habilidade, com maior ou menor grau de realismo, foram denominadas *chatbots*.

- Os produtos de NLP disponíveis no mercado podem ser descritos considerando outros sub-ramos para NLP, mas fazem isso para evidenciar funcionalidades específicas, as quais são diretamente deriváveis dos 3 sub-ramos acima.
- Segue um exemplo de um sistema Question & Answer, muito comum em assistentes de voz.



Fonte: https://developer.ibm.com/articles/a-beginners-guide-to-natural-language-processing/

- Ontologia, em NLP e em outras áreas da computação, se refere ao conhecimento de domínio (*domain knowledge*).
- Ontologia envolve taxonomia, atributos, relações, regras e eventos.

- "O Processamento de Linguagem Natural descreve um ramo da Inteligência Artificial que automatiza a compreensão e a síntese de linguagem para que os computadores e os seres humanos possam se comunicar sem intermediários. Para interagir com humanos, os computadores devem adotar e compreender sintaxe (gramática), semântica (significado da palavra), morfologia (e.g. passado, presente, futuro; gênero; número) e pragmática (efeitos do ato da conversação)."

Fonte: https://www.dataversity.net/what-is-natural-language-processing-nlp/

- São tarefas que se mostram bastante complexas, impondo muitos desafios para serem adequadamente assimiladas por uma máquina.
- Além disso, o NLP envolve técnicas multidisciplinares, fundamentadas em conceitos de linguística, ciência da computação, matemática, estatística, inteligência artificial, dentre outras áreas.

- A capacidade de comunicação requer que a intenção do falante seja compatível com a capacidade de percepção e compreensão do ouvinte, ambas vinculadas a um contexto em que a linguagem está sendo empregada. Aspectos como ambiguidade e finitude da linguagem criam enormes obstáculos para que a intenção do falante seja efetivamente capturada pelo ouvinte.

Fonte: https://www.shutterstock.com/pt/search/listening+reading+speaking+writing

- Para o desenvolvimento da Inteligência Artificial, é fundamental que as máquinas tenham capacidade de aprendizado e de comunicação a fim de permitir o aprendizado por experiência e interação, assim como acontece com os seres humanos (BARONI et al., 2017).

- As redes neurais artificiais, com destaque para as redes com arquitetura profunda, possibilitam isso através de múltiplos níveis de representação ao realizar a composição de camadas que aumentam o grau de abstração a cada nível, permitindo o aprendizado de funções mais complexas (LECUN et al., 2015).

- O domínio da linguagem humana pelas máquinas possibilita o acesso a uma parte significativa do conhecimento humano, pois este está codificado em linguagem natural. Além disso, cabe salientar que, de formas variadas, máquinas podem compartilhar seus cérebros, algo inacessível a seres humanos.

3 Corpus

- Segundo o maior linguista de *corpus* em língua inglesa de todos os tempos, John Sinclair, um *corpus* é uma coleção de textos de uma linguagem, em forma eletrônica, selecionada de acordo com critérios externos e capaz de representar, tão bem quanto possível, uma língua ou variedade linguística, de modo a servir como uma fonte de dados para pesquisas linguísticas.
- John Sinclair foi o editor-chefe do projeto que conduziu ao dicionário COBUILD (COLLINS Birmingham University International language Database), o primeiro a ser compilado a partir de um *corpus* digital, que hoje conta com 4,5 bilhões de palavras (https://collins.co.uk/pages/elt-cobuild-reference-the-collins-corpus).
- Embora a maioria dos *corpora* disponíveis contenham apenas texto, tem havido um número crescente de *corpora* multimodais, incluindo fala, sinais, postura corporal e gestos, os quais são selecionados de forma criteriosa, a partir de fontes autênticas da linguagem, e associados a meta-dados.

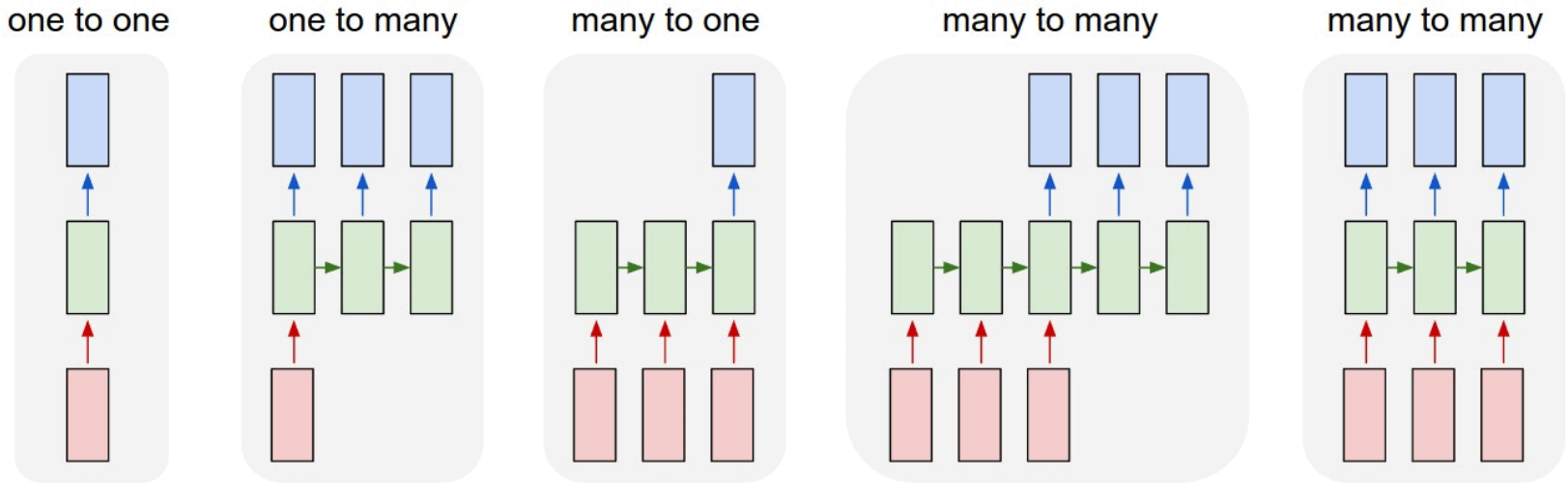
- Portanto, um *corpus* não se refere a uma coleção aleatória de amostras de uma linguagem, contemplando aspectos como representatividade e eliminação de vieses derivados da estratégia de coleta. Exemplos para a língua inglesa em: http://esl.fis.edu/learners/websites/corpora.htm
- Para mais detalhes, consultar:

  https://nordiskateckensprak.files.wordpress.com/2014/01/knb_whatisacorpus_cph-2013_outline.pdf

4 Uma nova visão para mapeamentos de entrada-saída

- Até aqui, neste curso, mesmo já tendo coberto neurocomputadores especializados em processar sequências, como as redes neurais recorrentes, não foi tomada ainda a iniciativa de enriquecer um pouco mais o cenário de mapeamentos de entrada-saída, diferenciando vetores de sequências de vetores.

- Pois é chegada a hora. A figura a seguir permite diferencias mapeamentos multidimensionais (a) vetor-para-vetor, (b) vetor-para-sequência, (c) sequência-para-vetor e (d),(e) sequência-para-sequência.



(a) (b) (c) (d) (e)

Fonte: https://cv-tricks.com/rnn/getting-started-recurrent-neural-networks/

5 Codificação densa para o tratamento de linguagem natural

- A representação *one-hot* para linguagem natural tem uma interpretação negativa e outra positiva:
  - ✓ O aspecto negativo é tornar todas as palavras de uma linguagem, vistas como símbolos atômicos, equidistantes entre si, impedindo qualquer representação semântica ou sintática. Além disso, trata-se de uma representação esparsa: por exemplo, numa linguagem com 2.500 palavras ou símbolos distintos, cada palavra ou símbolo da linguagem é representada por um vetor com 2.499 zeros e apenas 1 elemento unitário.
  - ✓ O aspecto positivo também tem a ver com a equidistância entre as palavras, pois essa condição pode ser explorada na síntese de mapeamentos capazes de estabelecer relações semânticas ou sintáticas entre as palavras ou símbolos da linguagem, no sentido, por exemplo, de mapear próximas entre si, no espaço de destino, palavras que operam como

sinônimos. É equivalente ao que se faz ao inicializar os pesos sinápticos de uma rede neural rasa: por não saber como o mapeamento a ser sintetizado deve se contorcer, a proposta é iniciar os pesos de modo a produzir algo próximo a um hiperplano, ou seja, um mapeamento sem contorção, que deve ser contorcido durante o treinamento da rede neural.

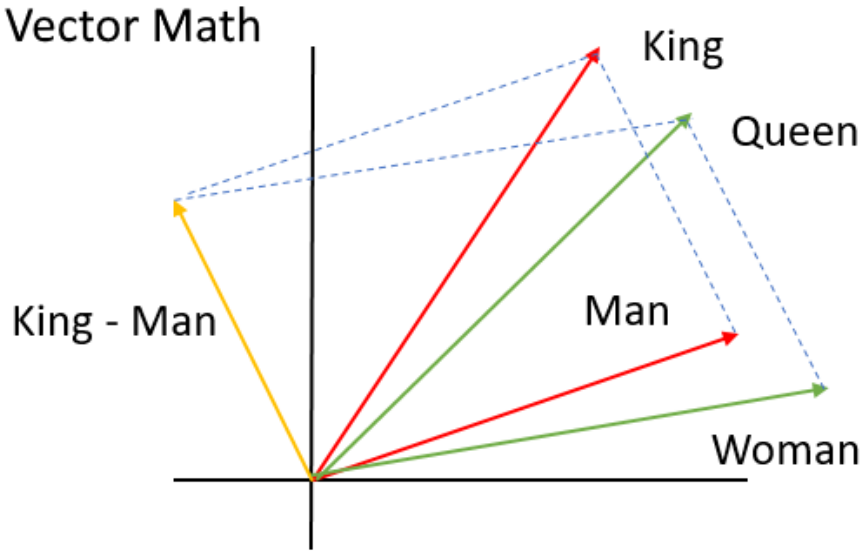
- A esparsidade da representação *one-hot* pode ser eliminada ao se propor uma redução de dimensão neste mapeamento, para um espaço de dimensão apropriada.
- Neste sentido, as soluções estado-da-arte em processamento de linguagem natural tendem a operar com um espaço denso de representação, promovendo *word embedding* ou codificação densa.
- Neste espaço de codificação densa, palavras semanticamente similares tendem a ser mapeadas próximas entre si. Além disso, a translação no espaço de codificação densa também pode codificar informação. Por exemplo, o deslocamento no espaço de codificação densa para se ir do vetor ‘king’ ao vetor ‘man’, por exemplo, tende

a ser bem semelhante ao deslocamento necessário para se ir do vetor ‘queen’ ao vetor ‘woman’. Alguns resultados de codificação permitiram até usar a composição para produzir novos significados: por exemplo, a soma do vetor ‘Germany’ e do vetor ‘capital’, produzem um resultado semelhante à representação de ‘Berlin’ (MIKOLOV et al., 2013a; 2013b).

- As ideias pioneiras para gerar *word embeddings* baseiam-se na Hipótese Distribucional do linguista inglês J.R. Firsth, que afirmava que itens com distribuições similares tendem a ter significados semelhantes. Isso é resumido na sua famosa frase: “Você conhecerá uma palavra pela sua companhia”.
- Os métodos distribucionais comprimem as informações de co-ocorrência de palavras em um dado contexto para obter uma representação. Um dos métodos mais relevantes é o *Latent Semantic Analysis* (LSA), que aplica decomposição em valores singulares em uma matriz contendo a contagem de palavras por documento (DEERWESTER et al., 1990; LANDAUER et al., 1998).

- Modelos neurais também são capazes de extrair relações entre palavras baseado somente na informação distribucional. Nesses modelos, o contexto usualmente são as palavras próximas em uma frase. Um dos principais modelos nessa linha é o Modelo Neural e Probabilístico de Linguagem (*Neural Probabilistic Language Model*, NPLM) (BENGIO et al., 2003), que utiliza uma rede neural a fim de prever a próxima palavra, dado um contexto de tamanho fixo. Foi um dos pioneiros no uso de redes neurais para NLP com *word embedding* e motivou arquiteturas posteriores, que superaram em desempenho técnicas baseadas em LSA.
- Já deve ter ficado claro, portanto, que NLP requer um codificador: mapeamento que recebe como entrada um vetor *one-hot* (ou um vetor com alguns poucos elementos não-nulos em aplicações que operam com múltiplas palavras de entrada) e produz na saída um ponto num espaço de características, também denominado de espaço de codificação densa ou *word embedding*.

- Este codificador pode ser bastante complexo, o que justifica a adoção de *deep learning* em processamento de linguagem natural. Mas dependendo da aplicação e buscando escalabilidade, também já foram considerados mapeamentos lineares.
- Uma dimensão elevada para o espaço de codificação densa torna mapeamentos lineares suficientemente poderosos em algumas aplicações, permitindo realizar operações semânticas na linguagem com operadores aritméticos no espaço de codificação densa.



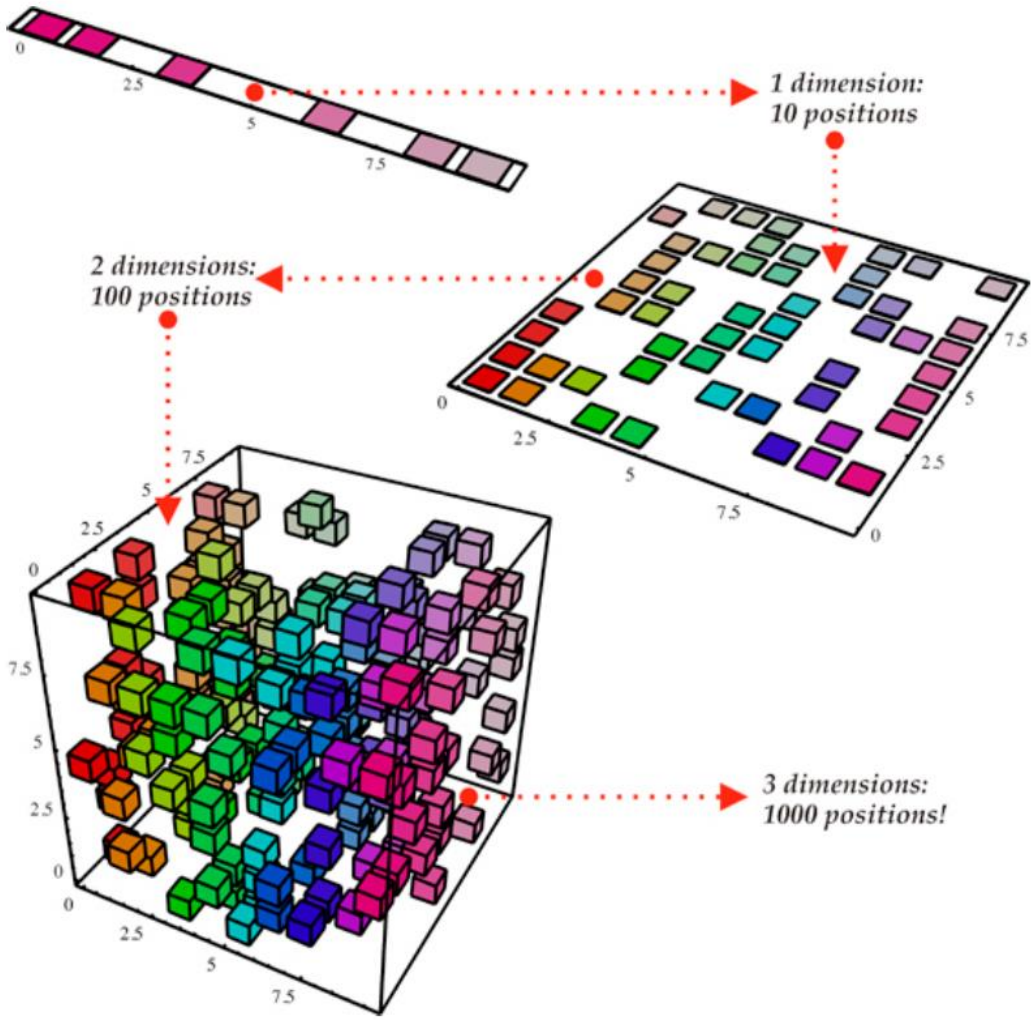


Figura 1 – Efeito do aumento de dimensão num cenário em que cada coordenada admite10 “símbolos ordenados” possíveis

- Cabe observar que, em espaços de codificação densa, diferente da codificação *one-hot*, passa a existir uma relação de ordem entre os valores de cada coordenada e, com isso, é possível voltar a falar de distância, similaridade e vizinhança entre pontos.
- Mais ainda, com o aumento no número de coordenadas, as possibilidades de múltiplos objetos, neste espaço de codificação densa, estabelecerem relações de vizinhança aumentam exponencialmente.

6 Word2vec

- Uma proposta de mapeamento linear é o word2vec (MIKOLOV et al., 2013a) e suas adaptações e extensões, como em MIKOLOV et al. (2013b), que se mostram bem escaláveis. Foram utilizados bilhões de registros para se treinar uma estrutura do tipo codificador-decodificador, sendo que o gargalo tem dimensão de 50 a 300 unidades.

- As regularidades construídas automaticamente no espaço de codificação densa produzido levaram a associações do tipo: o vetor ‘Rome’ é o resultado da operação entre vetores: ‘Paris’ – ‘France’ + ‘Italy’.
- Essencialmente o word2vec é uma rede neural MLP com apenas uma camada intermediária e função de ativação identidade para todos os neurônios. A operação linguística de entrada-saída que a rede neural aprende pode ser variada. Exemplos são previsão de próxima palavra e de palavra central. O sucesso na realização dessas tarefas linguísticas requer a concepção de um espaço de codificação densa que carrega sintaxe, semântica, morfologia e pragmática da linguagem.
- As figuras a seguir ilustram a operação. Quando aparecem múltiplos vetores de entrada, como no modelo *Continuous Bag of Words* (CBOW), ou múltiplos vetores de saída, como no modelo *Skip-Gram*, a matriz que aparece em multiplicidade é a mesma.

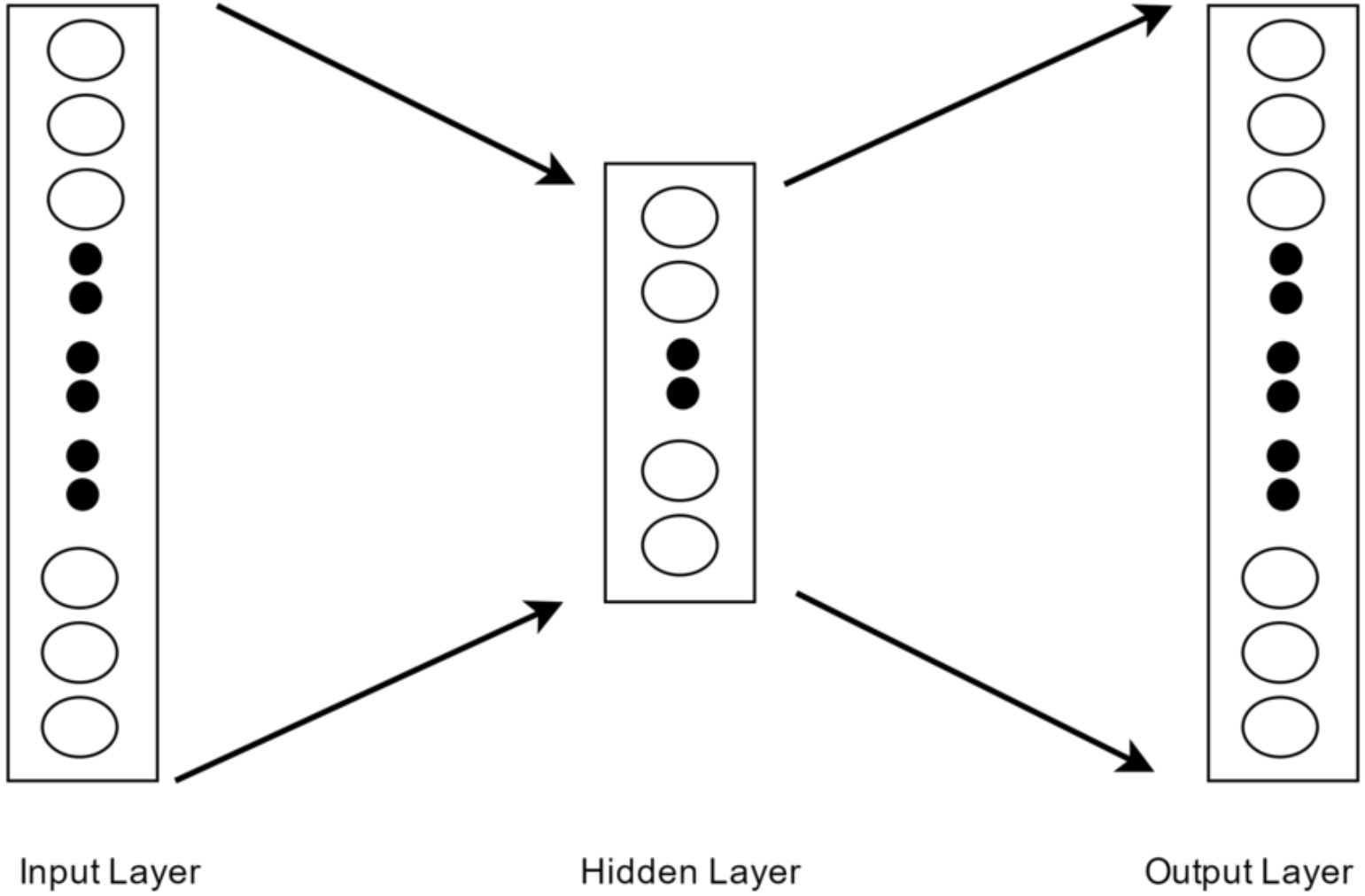


Figura 2 – Arquitetura adotada para o word2vec, com neurônios apresentando funções de ativação lineares

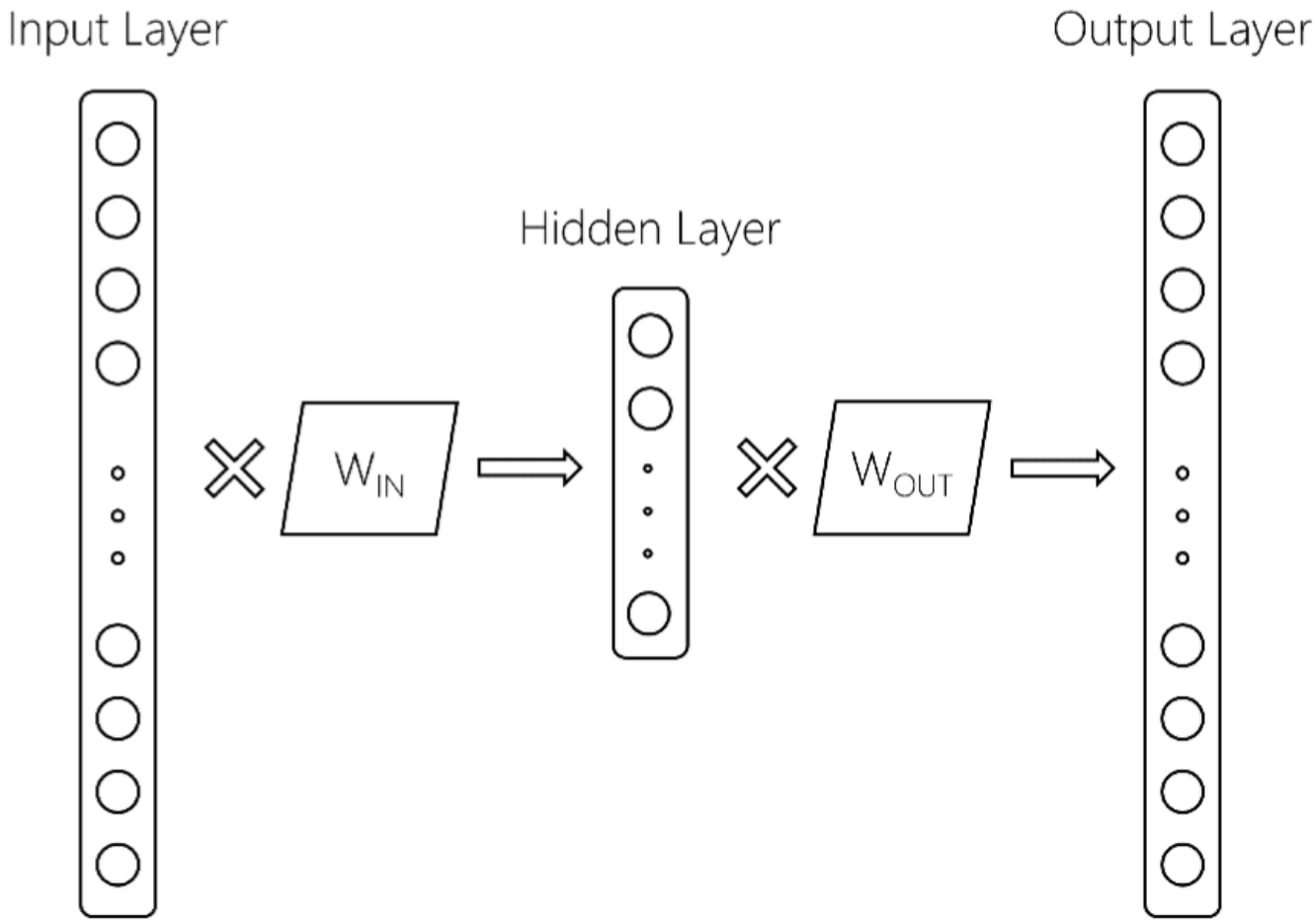


Figura 3 – Outra perspectiva da mesma arquitetura de rede para o word2vec. As matrizes $\mathbf{W_{IN}}$ e $\mathbf{W_{OUT}}$ podem ser forçadas a serem transpostas entre si, dependendo da aplicação.

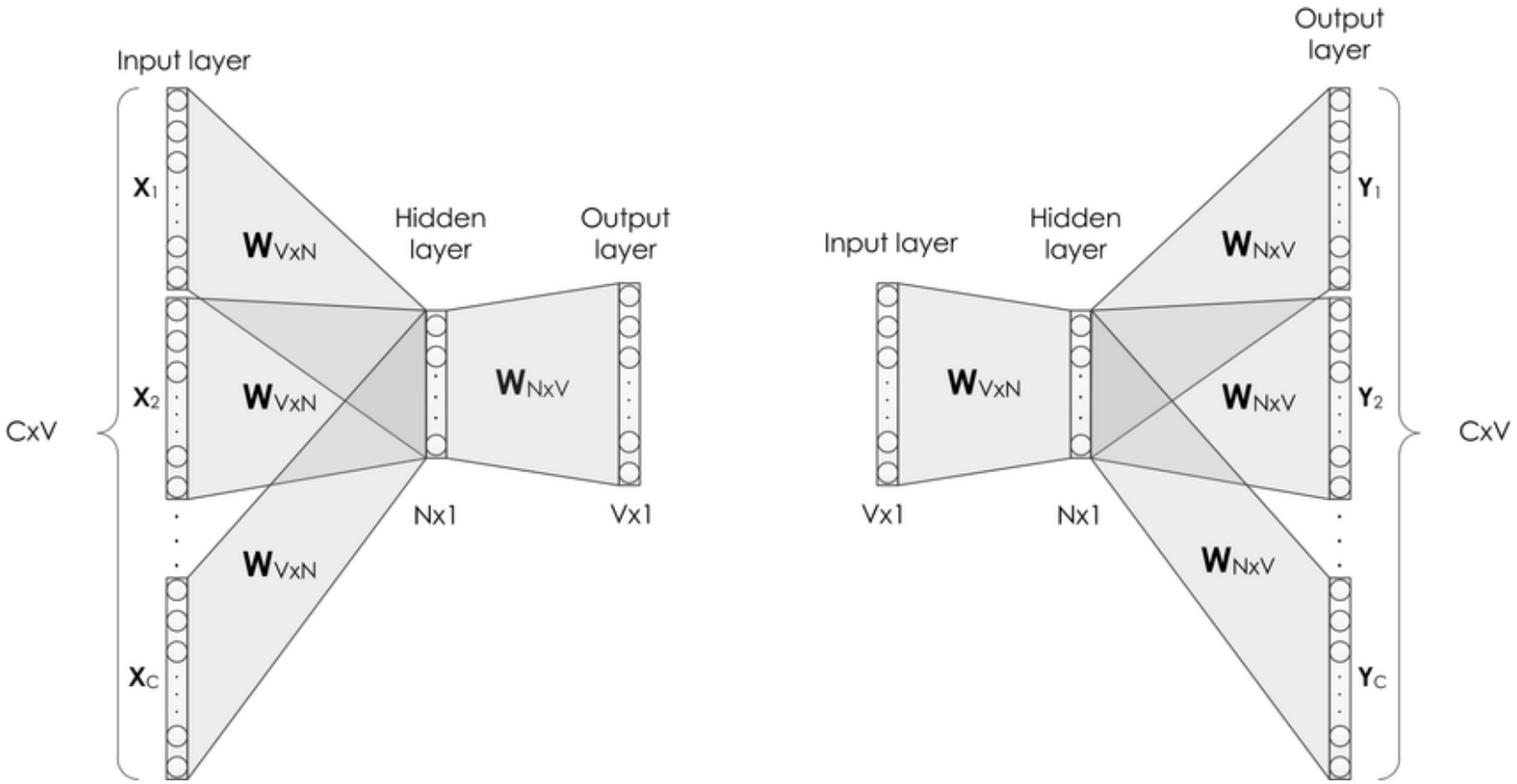


Figura 4 – À esquerda se tem o modelo CBOW e à direita o Skip-Gram.

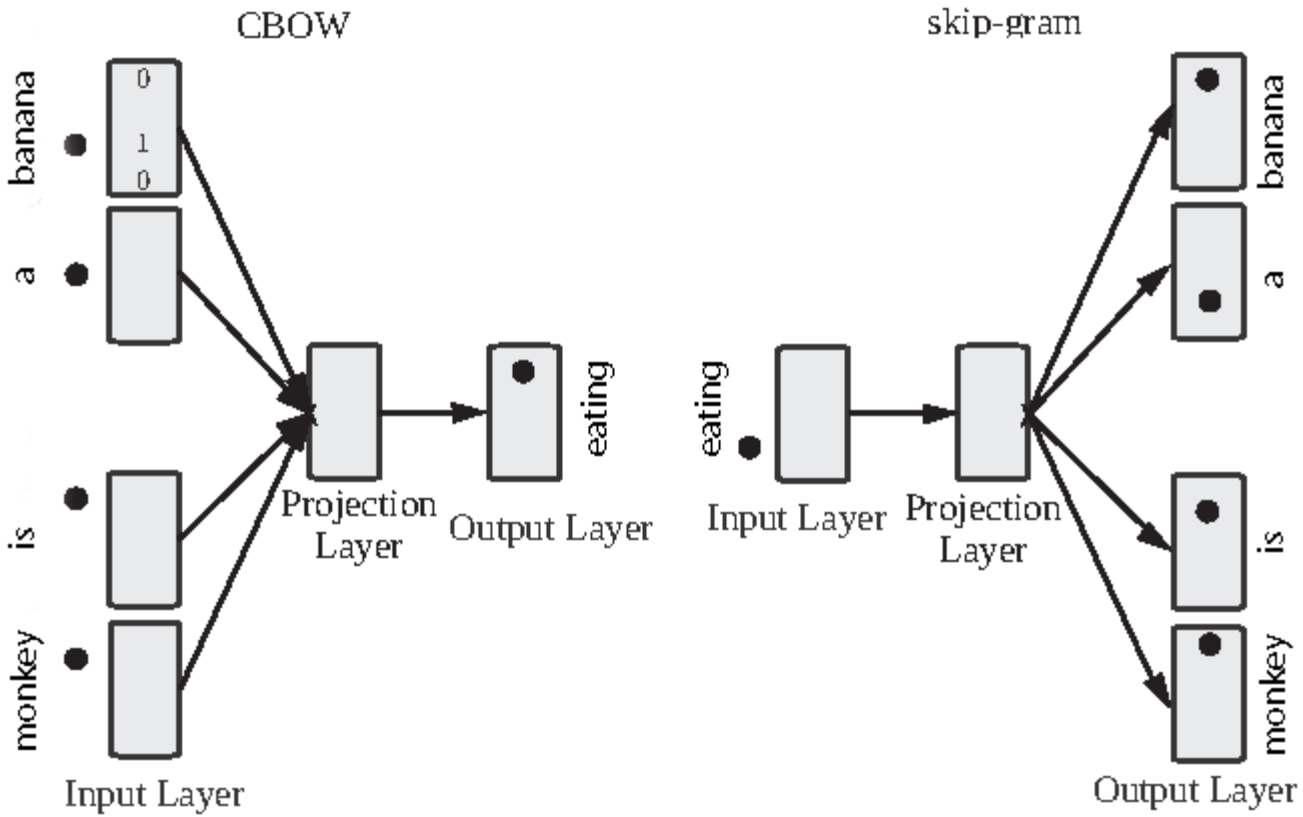


Figura 5 – Exemplo de resultado produzido por CBOW (à esquerda) e Skip-Gram (à direita).

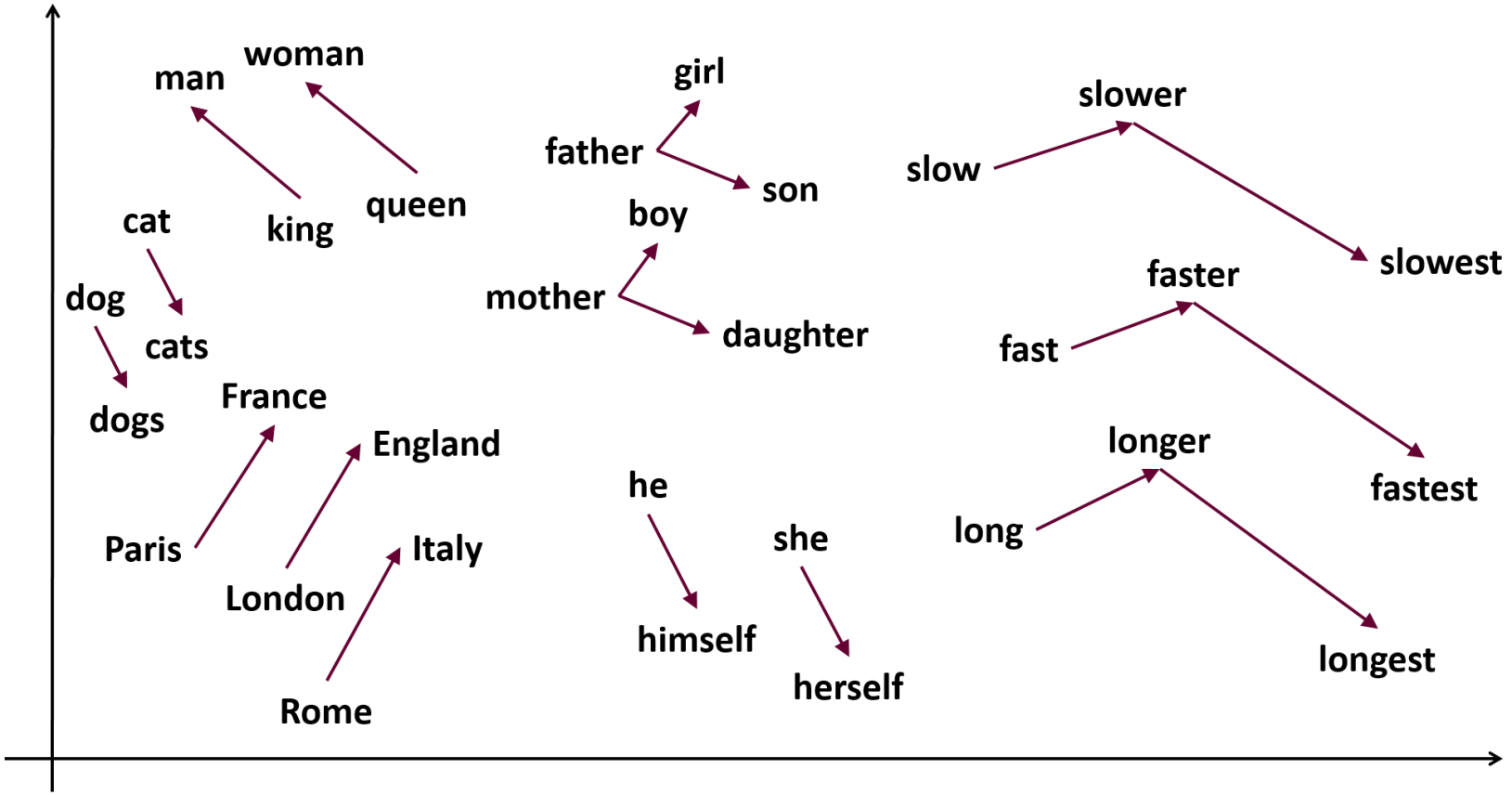


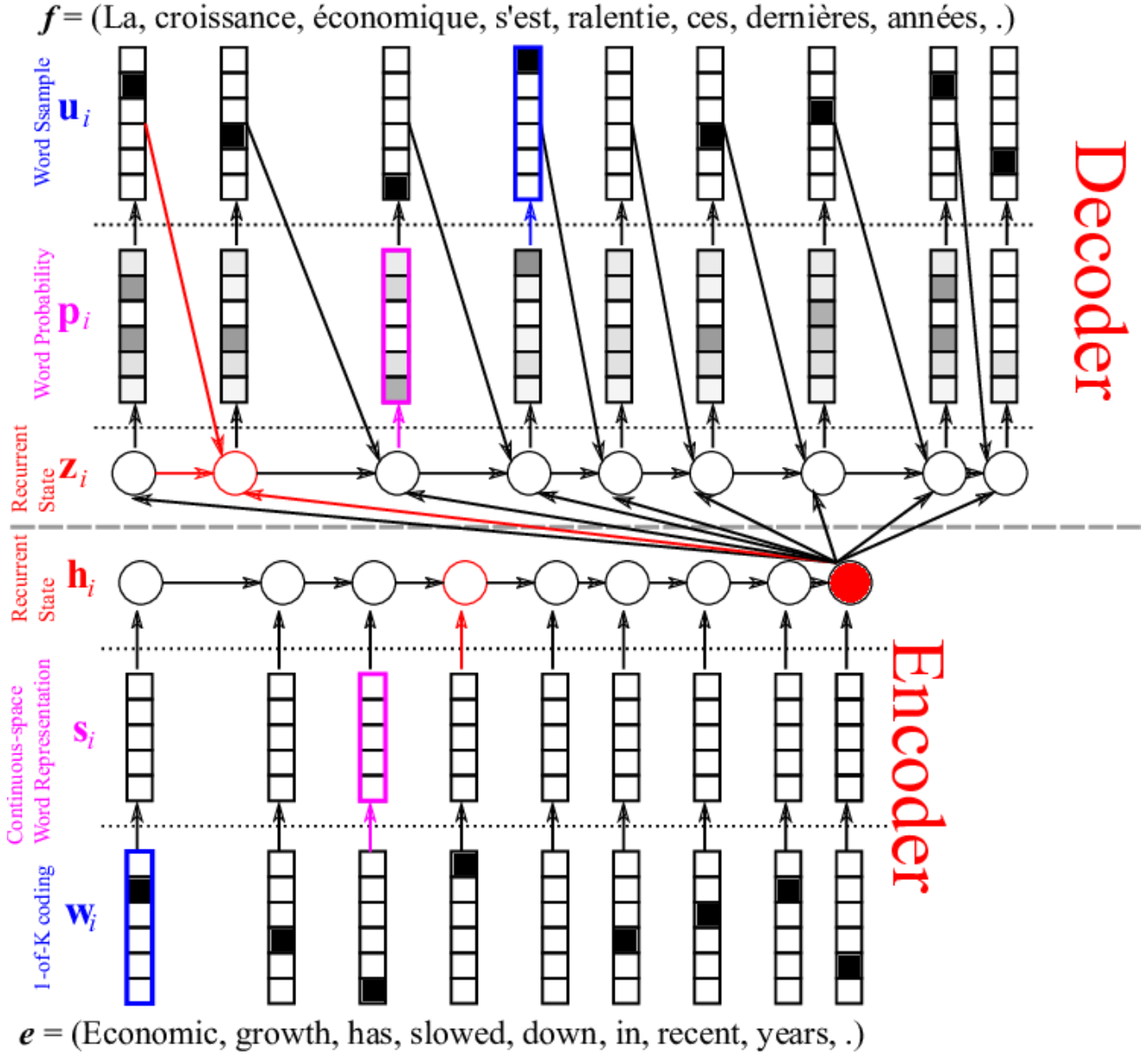
Figura 6 – Projeção 2D de alguns vetores do espaço de codificação densa produzido por modelos word2vec.

7 Máquinas tradutoras de frases

- Esta seção contém material extraído de: https://devblogs.nvidia.com/introduction-neural-machine-translation-with-gpus/
- A parte conceitual está predominantemente fundamentada em CHO et al. (2014).
- O modelo a ser adotado aqui é o de codificador-decodificador, contendo dois módulos recorrentes. A entrada é uma frase numa linguagem e a saída é uma tradução para a frase de entrada, numa outra linguagem. Um elenco de pares de frases deve ser utilizado para guiar o processo de treinamento deste modelo.
- O primeiro módulo recorrente, na saída do codificador, é responsável por sumarizar sequências de entrada de comprimento variável, já convertidas de *one-hot* para uma codificação densa. Essa sumarização se dá na forma de um vetor de comprimento fixo, que nada mais é do que o vetor de estados deste primeiro módulo recorrente. Este vetor de estados é atualizado sequencialmente, com cada palavra da frase de entrada, em sua codificação densa.

- Já o segundo módulo recorrente é o primeiro elemento do decodificador. Com base em (1) seu estado interno, (2) última palavra definida para a frase de saída (quando houver) e (3) saída do codificador, este módulo recorrente propõe sequencialmente vetores de probabilidade, que guiam o processo de amostragem da sequência de símbolos de saída. Portanto, a saída do decodificador será a frase traduzida, também de comprimento variável.
- O critério de desempenho é maximizar a probabilidade condicional da frase esperada na saída (sequência-alvo) da máquina tradutora, dada a proposta de frase de entrada (sequência-fonte), considerando todo o conjunto de treinamento.
- Não será dada ênfase aqui às estratégias de treinamento do modelo, mas sim à sua funcionalidade e à estrutura do modelo de aprendizado. Mas cabe comentar que codificador e decodificador serão treinados conjuntamente para maximizar a probabilidade da frase de saída (sequência-alvo) condicionada à ocorrência da frase de entrada (sequência-fonte), considerando todos os pares de frases.

- A rede neural, portanto, vai operar como uma máquina de tradução estatística: a frase de saída é amostrada sequencialmente, palavra a palavra, a partir da obtenção de um vetor de probabilidades que indica a chance de escolha de cada palavra candidata, o qual é usado para amostrar a palavra seguinte da sequência de saída.
- Uma análise deste modelo codificador-decodificador, realizada por CHO et al. (2014), revela que ele sintetiza um espaço de codificação densa capaz de preservar a informação linguística contida na sequência-fonte.
- A figura a seguir descreve graficamente todas as etapas associadas ao codificador e ao decodificador, conforme a descrição textual acima. Os módulos recorrentes aparecem desdobrados no tempo, da esquerda para a direita, indicando que a chegada de palavras e a saída de palavras é sequencial.
- Em CHO et al. (2014), cada módulo recorrente é uma GRU (*Gated Recurrent Unit*), que é um bloco LSTM descrito na Parte 3 deste Tópico 8.


f = (La, croissance, économique, s'est, ralentie, ces, dernières, années, .)
Word Sample
u_i
Word Probability
p_i
Recurrent State
z_i
Decoder
Recurrent State
h_i
Continuous-space Word Representation
s_i
1-of-K coding
w_i
Encoder
e = (Economic, growth, has, slowed, down, in, recent, years, .)

7.1 Máquina de tradução estatística

- A GRU do codificador realiza a operação:

$$h_{\langle i \rangle} = f\left(h_{\langle i-1 \rangle}, s_{\langle i-1 \rangle}\right).$$

- O vetor de estados $h_{\langle i \rangle}$ usa codificação densa, ou seja, $h_{\langle i \rangle} \in \Re^M$, com M sendo uma dimensão apropriada, dependente de cada aplicação.
- Já a GRU do decodificador realiza a operação:

$$\begin{cases} z_{\langle i \rangle} = f\left(z_{\langle i-1 \rangle}, u_{\langle i-1 \rangle}, h_{\langle \text{final} \rangle}\right) \\ p_{\langle i \rangle} = \text{softmax}\left(z_{\langle i \rangle}\right) \end{cases}.$$

- Seja T o número de palavras da linguagem de saída, então o vetor $p_{\langle i \rangle}$ tem dimensão T, assim como o vetor $z_{\langle i \rangle}$, e cada elemento de $p_{\langle i \rangle}$ é dado na forma:

$$p_{\langle i \rangle, j} = \frac{\exp\left(z_{\langle i \rangle, j}\right)}{\sum_{l=1}^{T} \exp\left(z_{\langle i \rangle, l}\right)}.$$

- A rede neural codificadora-decodificadora, contendo dois módulos LSTM, é uma máquina de tradução estatística, pois busca maximizar a probabilidade de cada frase desejada para a saída (sequência-alvo), associada a cada frase de entrada apresentada durante o treinamento (sequência-fonte).
- Seja $\left\{w_1^{(q)}, w_2^{(q)}, \cdots, w_V^{(q)}\right\}$ a sequência de palavras (símbolos) de entrada para o *q*-ésimo par de frases de treinamento, lembrando que o número *V* de palavras da frase de entrada é variável (pode ser diferente para cada *q*).
- Seja $\left\{u_{1,j_1}^{(q)}, u_{2,j_2}^{(q)}, \cdots, u_{T,j_T}^{(q)}\right\}$ a sequência de palavras (símbolos) de saída para o *q*-ésimo par de frases de treinamento, lembrando que o número *T* de palavras da frase de saída é variável (pode ser diferente para cada *q*).
- A probabilidade de cada palavra desejada para a saída é, então, dada na forma:

$$p_{\langle i\rangle, j_i}^{(q)} = p\left(u_{i,j_i}^{(q)} \middle| w_1^{(q)}, w_2^{(q)}, \cdots, w_V^{(q)}, u_{i-1}^{(q)}, u_{i-2}^{(q)}, u_1^{(q)}\right),$$

sendo j_i o índice da *i*-ésima palavra no dicionário de todas as palavras possíveis.

- E a probabilidade conjunta da frase de saída assume então a forma:

$$\prod_{i=1}^{T} p^{(q)}_{\langle i \rangle, j_i} = \prod_{i=1}^{T} p\left(u^{(q)}_{i,j_i} \middle| w_1^{(q)}, w_2^{(q)}, \cdots, w_V^{(q)}, u^{(q)}_{i-1}, u^{(q)}_{i-2}, u_1^{(q)}\right).$$

- Aplicando o logaritmo à expressão acima e somando para todo q, o ajuste dos pesos da rede neural codificadora-decodificadora, contendo dois módulos LSTM, deve maximizar o seguinte critério de desempenho:

$$\sum_{q=1}^{N}\sum_{i=1}^{T} \log\left(p^{(q)}_{\langle i \rangle, j_i} \right) = \sum_{q=1}^{N}\sum_{i=1}^{T} \log\left[p\left(u^{(q)}_{i,j_i} \middle| w_1^{(q)}, w_2^{(q)}, \cdots, w_V^{(q)}, u^{(q)}_{i-1}, u^{(q)}_{i-2}, u_1^{(q)}\right)\right],$$

  lembrando que T e V variam com q.

7.2 O papel semântico do codificador

- Da mesma forma como foi verificado no caso do word2vec, o codificador da máquina tradutora de frases tem um papel semântico bem estabelecido. A figura a seguir apresenta projeções 2D do vetor de saída do codificador para algumas frases presentes na entrada, após o treinamento.

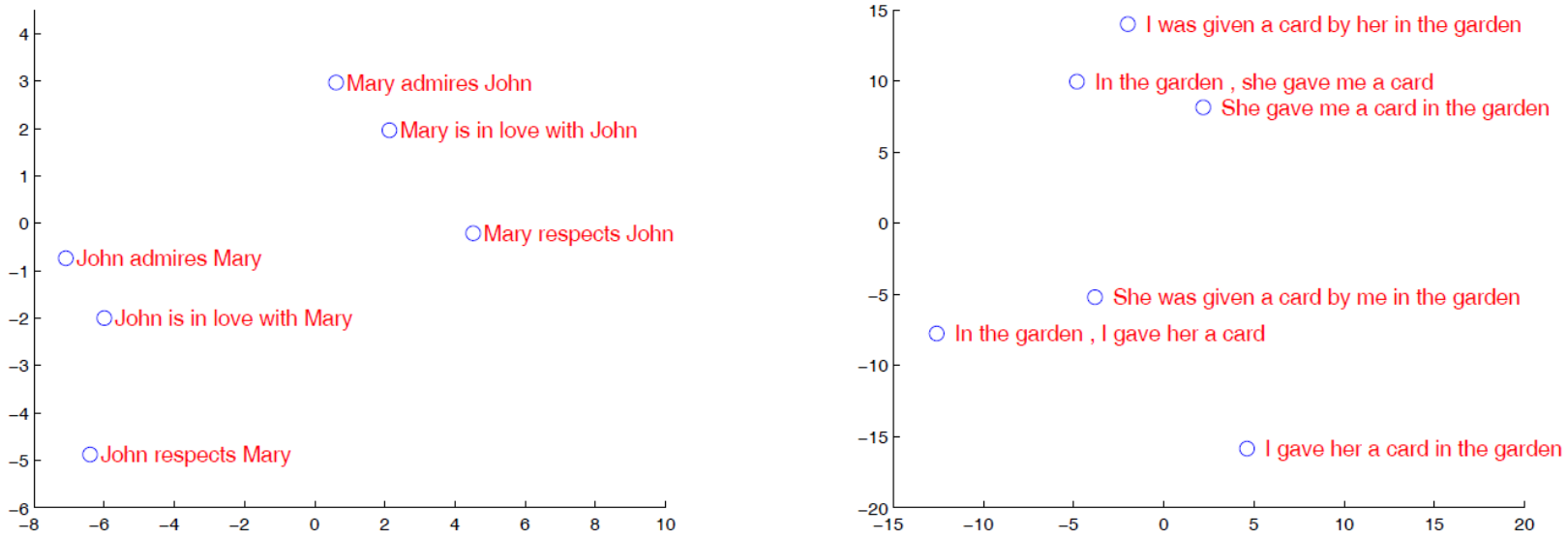
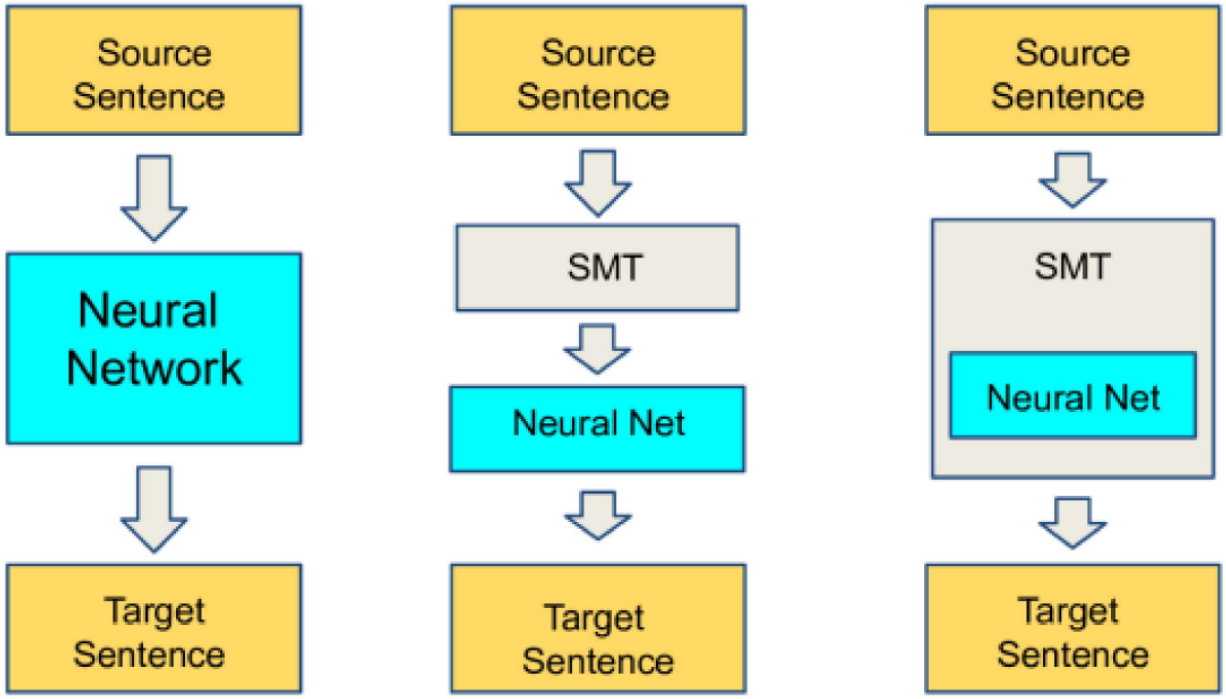


Figura 7 – Comportamento do vetor de sumarização (saída do codificador) para algumas frases de entrada (figura extraída de SUTSKEVER et al. (2014)).

- Outros aspectos linguísticos, além da questão semântica, também podem ser contemplados neste vetor de sumarização.
- Em virtude do tratamento sequencial das frases e seu mapeamento num mesmo espaço de codificação densa, é possível comparar frases de tamanho e estrutura distintos. Trata-se de um resultado muito expressivo.

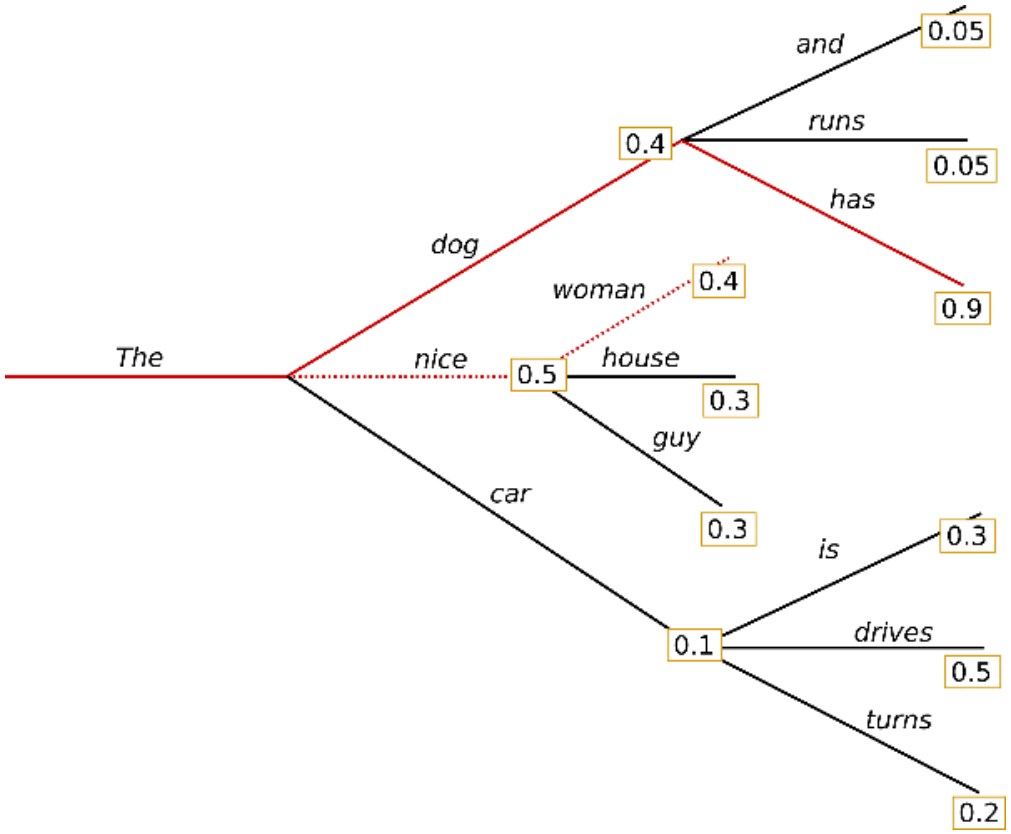


(a) (b) (c)

- Uma rede neural passou a cuidar sozinha de toda a máquina de tradução (a), após aparecer como participante do processo (b e c) em metodologias propostas anteriormente e que também chegaram a ser estado-da-arte.

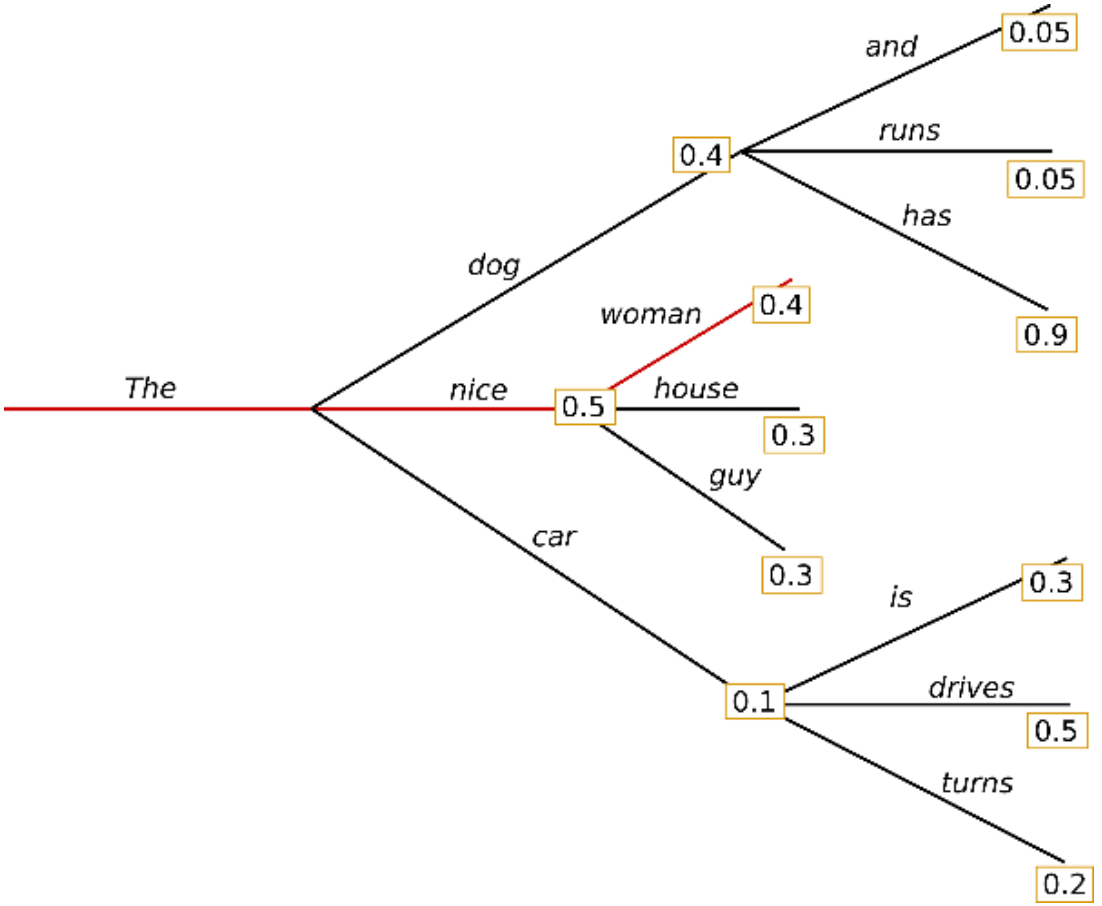
7.3 Beam search

- Uma vez treinada a máquina de tradução estatística, visando maximizar a chance de gerar a frase mais provável na saída, é possível empregar *beam search*.



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

- Essa técnica permite escapar de mínimos locais produzidos pela busca *greedy*, ilustrada a seguir:



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

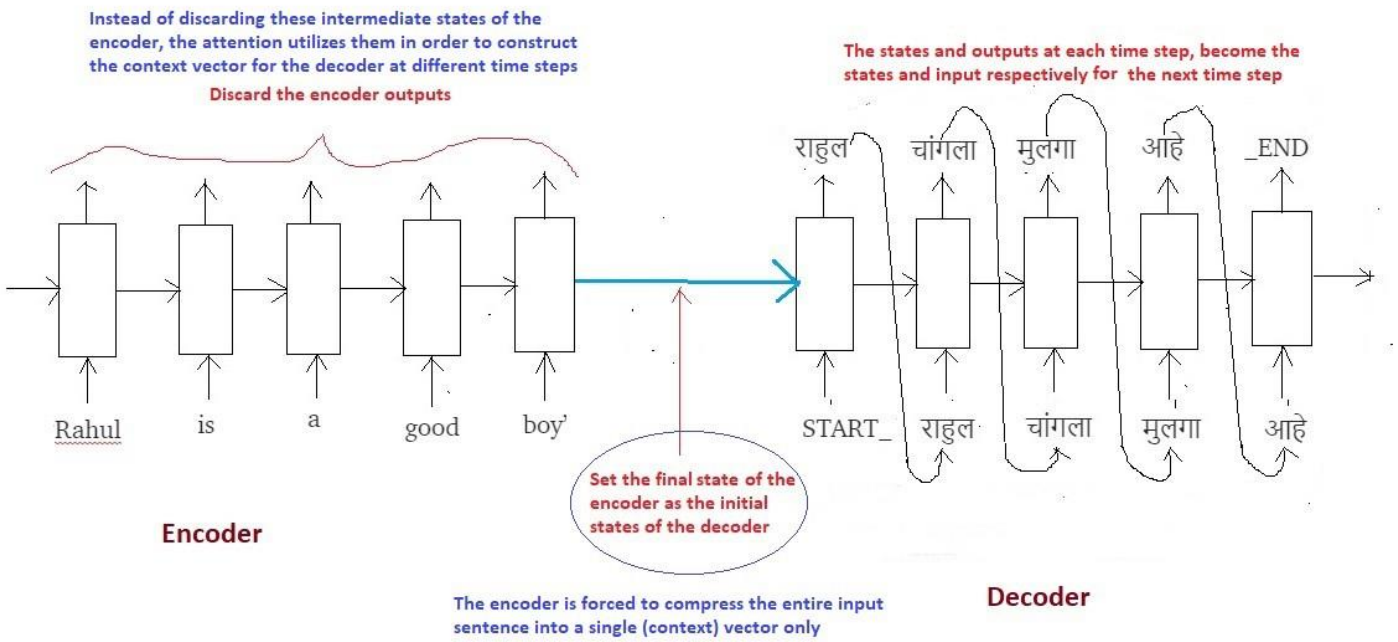
7.4 O índice BLEU para avaliação da tradução realizada

- BLEU (*BiLingual Evaluation Understudy*) (PAPINENI et al., 2002) é um índice que avalia a qualidade da tradução realizada por uma máquina, de uma linguagem natural para outra. O índice varia no intervalo [0,1] e é tão maior quanto mais próxima estiver a tradução feita pela máquina de uma tradução realizada por um especialista humano.
- O índice é uma composição das notas atribuídas a segmentos traduzidos, geralmente sentenças, pela comparação direta com um conjunto de traduções de referência para cada segmento. Todo o conjunto de segmentos do *corpus* deve ser considerado para se chegar à qualidade do resultado de tradução. Repare que inteligibilidade e correção gramatical não são diretamente ponderados.
- Cabe salientar que mesmo traduções feitas por especialistas humanos, distintos daqueles que geraram as sentenças de referência, dificilmente obtêm avaliação máxima pelo BLEU, pois muitas traduções possíveis não são consideradas.

8 Modelos de aprendizado com mecanismos de atenção

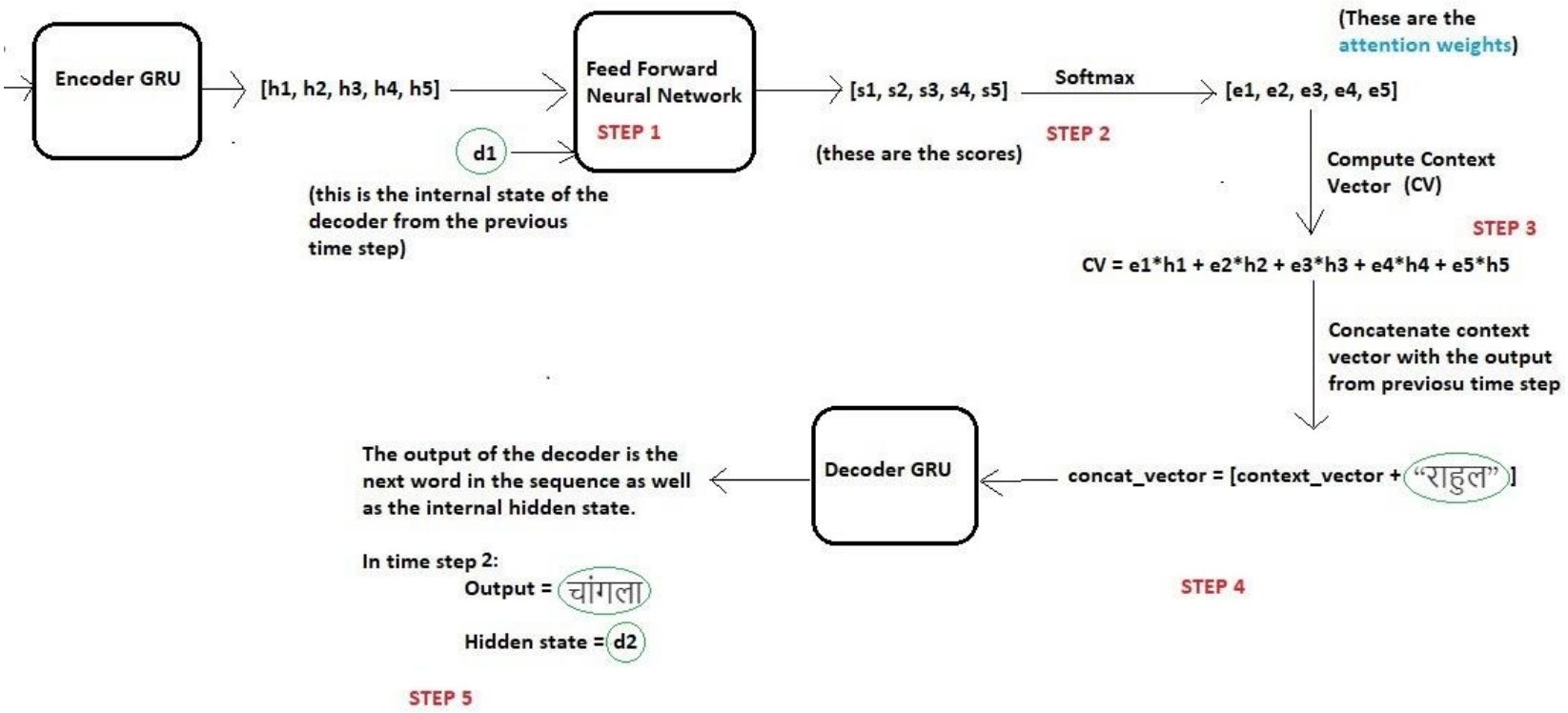
- Vamos abordar brevemente aqui uma das ideias de maior potencial de aplicação em *deep learning*: aprendizado com mecanismos de atenção (*attention learning*) (BAHDANAU et al., 2015). Dentre aplicações de destaque, cabe mencionar tradução de texto e rotulação de imagens.
- Os modelos denominados Seq2Seq, tendo a máquina tradutora de frases da seção 7 como um exemplo típico, foram os primeiros a se beneficiarem da incorporação de mecanismos de atenção. O ganho de desempenho foi tão expressivo que os tradutores estado-da-arte hoje operam apenas com eles (buscando escalabilidade, além de alto desempenho) ou tendo esses mecanismos como módulos principais.
- Será mantida aqui a aplicação envolvendo a síntese de uma máquina tradutora de texto, sendo o material desta seção predominantemente fundamentado em: https://towardsdatascience.com/intuitive-understanding-of-attention-mechanism-in-deep-learning-6c9482aecf4f

- A arquitetura de modelos Seq2Seq, como visto na seção 7, é composta de uma estrutura codificador-decodificador. O codificador sumariza a informação de entrada (que pode ser de tamanho variável) no estado interno (que é de tamanho fixo) do bloco LSTM, também denominado de vetor de contexto. Para o sucesso da decodificação, é esperado que este vetor de contexto represente um bom sumário da sequência de entrada inteira. O decodificador, então, é inicializado com este vetor de contexto para dar partida à síntese da sequência de saída.
- Uma limitação em potencial desta proposta centrada num único vetor de contexto de tamanho fixo é que a tarefa de codificação/decodificação tende a sofrer degradação de desempenho com um aumento de tamanho da sequência de entrada.
- Os mecanismos de atenção foram justamente propostos para lidar com esta limitação. A figura a seguir ilustra a informação que é descartada quando não se adota um mecanismo de atenção.

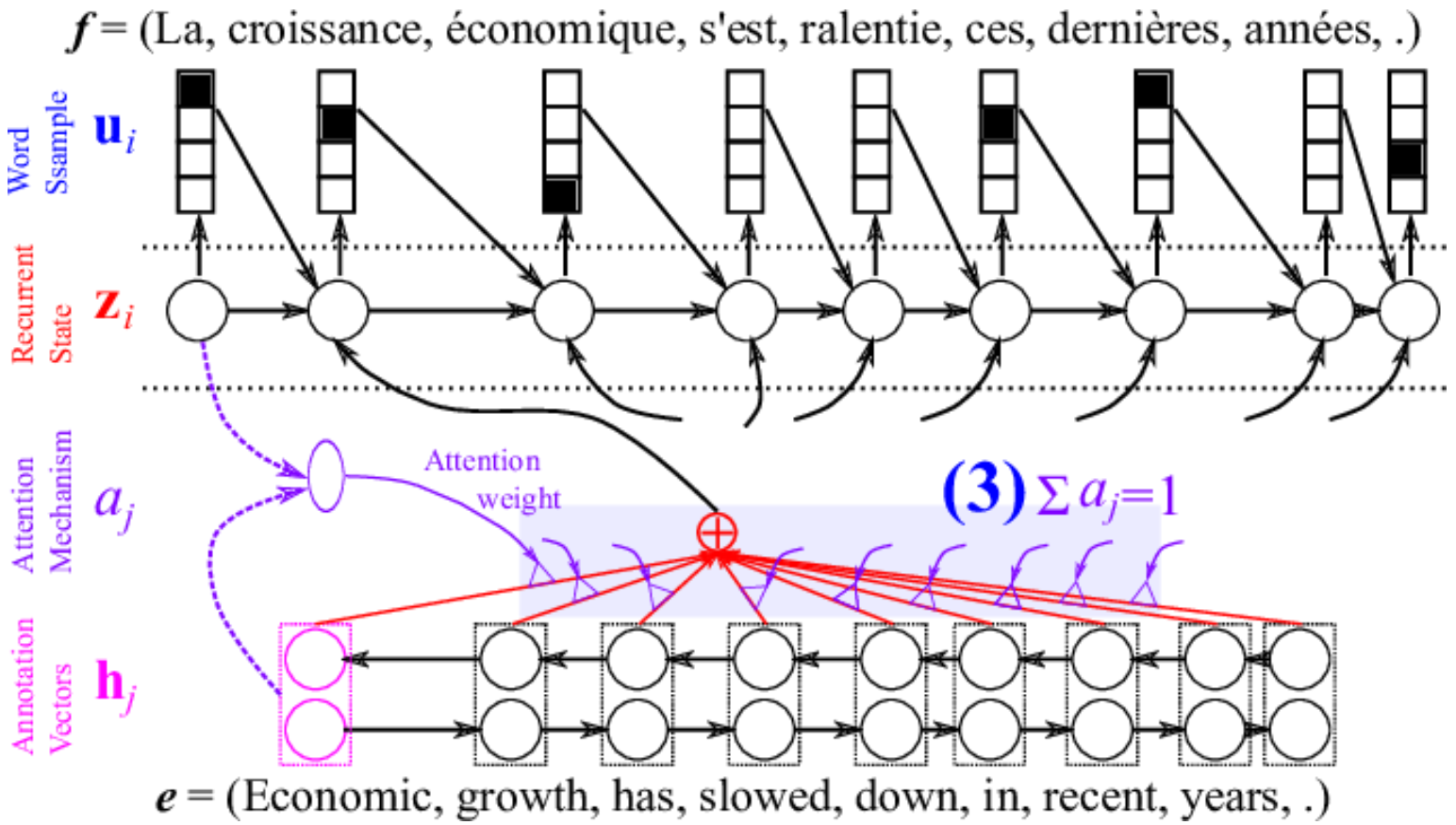


- Todos os estados intermediários do codificador são descartados, sendo utilizado apenas o seu estado final para inicializar o decodificador. Na prática, esta estratégia se mostrou eficiente para sequências curtas e uma limitação para sequências mais longas.
- A ideia central dos mecanismos de atenção é utilizar ponderações de todos esses estados intermediários na construção do vetor de contexto.

- Repare que esses estados intermediários são vetores de mesma dimensão e que armazenam informação local acerca da sequência de entrada.
- Evidentemente, não é uma tarefa elementar definir ponderações adequadas a cada instante de operação, ou seja, definir em que informações locais a atenção deve ser focada a cada instante.
- Um mecanismo de atenção competente é indiretamente obtido sempre que o comportamento de entrada-saída da máquina de aprendizado for adequadamente capturado, após o ajuste de pesos por retropropagação do erro. Em outras palavras, chega-se a um bom mecanismo de atenção ao minimizar o erro na saída da máquina de aprendizado.
- Repare que a função softmax é utilizada para a definição dos pesos de ponderação associados ao mecanismo de atenção, conforme esquematizado na figura da próxima página, onde o codificador recebe a sequência de 5 palavras “Rahul is a good boy”.



- Além de um ganho de desempenho, a introdução de mecanismos de atenção amplia a interpretabilidade dos modelos Seq2Seq, no sentido de explicar o que foi relevante, na entrada, para a definição da saída no instante corrente.
- Para mais detalhes: https://www.youtube.com/watch?v=SysgYptB198 e https://www.youtube.com/watch?v=quoGRI-1l0A

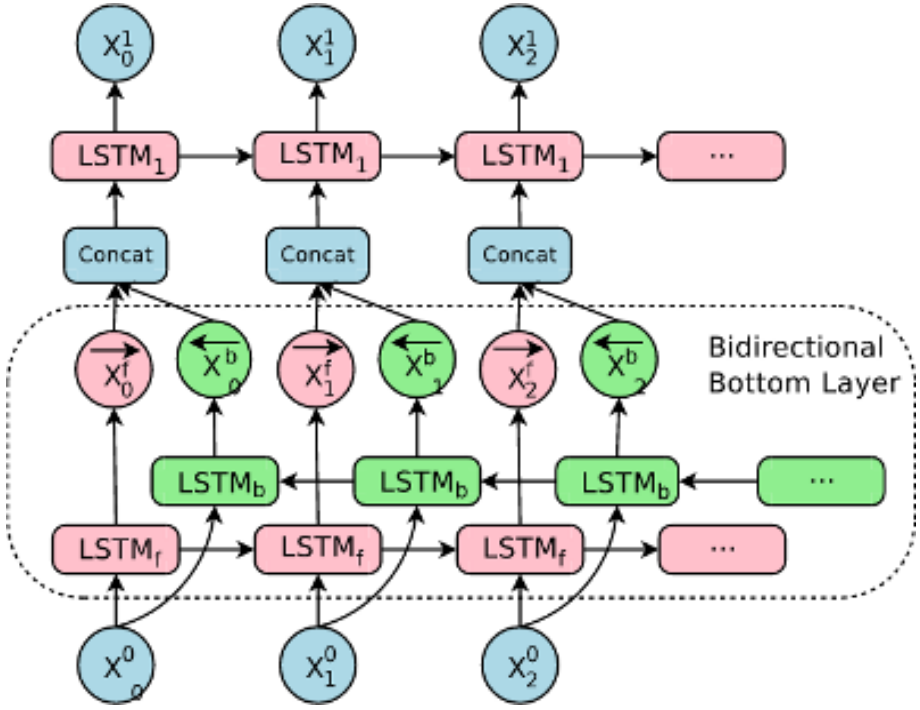


Fonte: https://devblogs.nvidia.com/introduction-neural-machine-translation-gpus-part-3/

- Há também modelos de atenção aditivos e multiplicativos (GÉRON, 2019).

9 Codificador bidirecional

- O processamento de qualquer sequência (para efeito de tradução, por exemplo) pode depender de eventos passados e também de eventos futuros da sequência. Uma adaptação simples aos modelos Seq2Seq apresentados até aqui é a adoção de um codificador bidirecional, como ilustrado na figura a seguir.



Fonte: https://arxiv.org/pdf/1609.08144v2.pdf

10 Um pouco de história em NLP

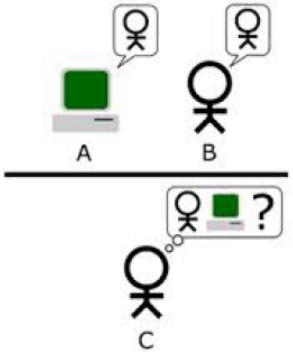
- As metodologias estado-da-arte para NLP, ao longo das últimas décadas, podem ser sintetizadas na seguinte sequência:
  - mid-1970s: **HMMs** for speech recognition ⇒ probabilistic models
  - early 2000s: **conditional random fields** for part-of-speech tagging ⇒ structured prediction
  - early 2000s: **Latent Dirichlet Allocation** for modeling text documents ⇒ topic modeling
  - mid 2010s: **sequence-to-sequence models** for machine translation ⇒ neural networks with memory/state

Slide extraído de Liang, P. "Natural Language Understanding: Foundations and State-of-the-Art", ICML Tutorial, 2015.

11 O jogo da imitação

The Imitation Game (1950)

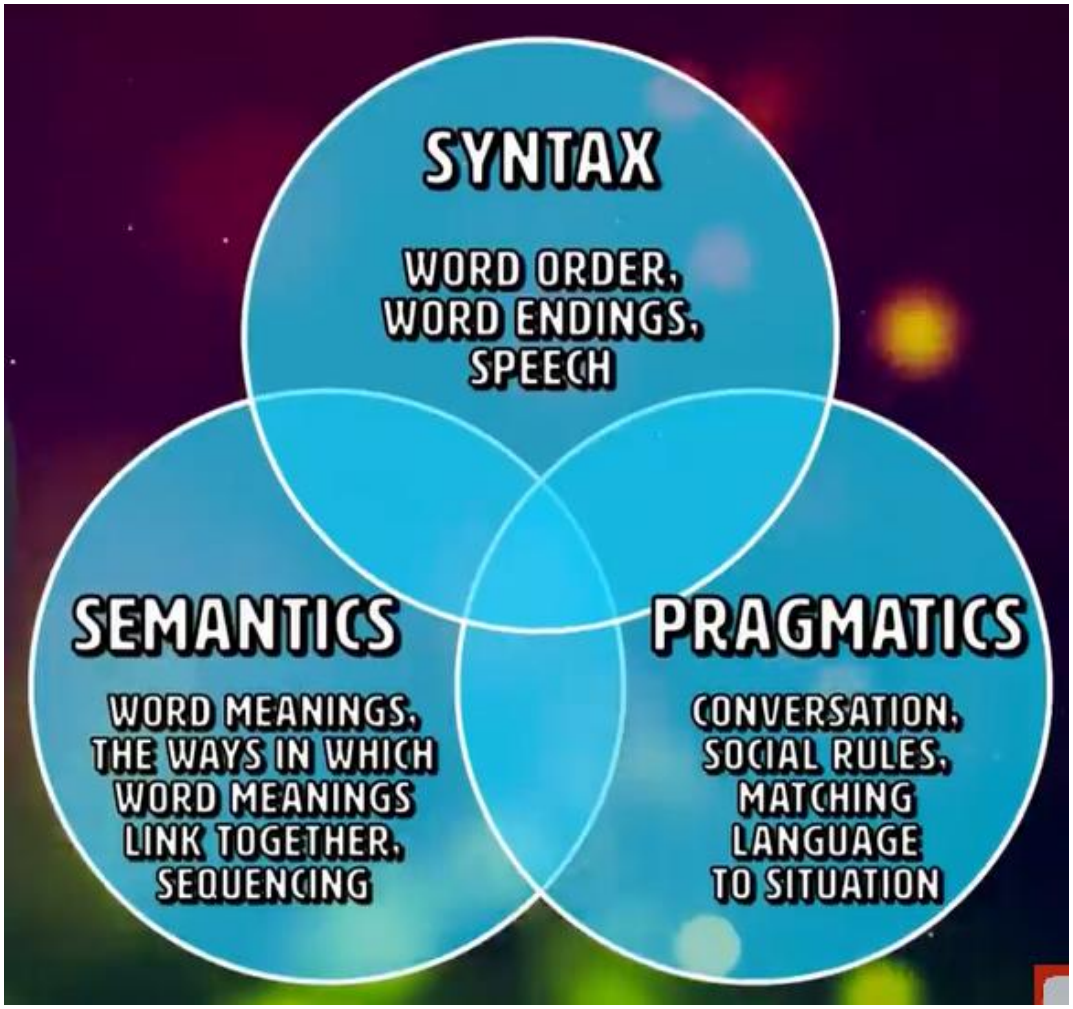
"Can machines think?"



Q: Please write me a sonnet on the subject of the Forth Bridge.
A: Count me out on this one. I never could write poetry.
Q: Add 34957 to 70764.
A: (Pause about 30 seconds and then give as answer) 105621.

- **Behavioral** test
- ...of **intelligence**, not just natural language understanding

Slide extraído de Liang, P. "Natural Language Understanding: Foundations and State-of-the-Art", ICML Tutorial, 2015.



Fonte: https://www.youtube.com/watch?v=bDxFvr1gpSU

12 Análise de sentimento

Sentiment Detection and Bag-of-Words Models

- Sentiment is that sentiment is "easy"
- Detection accuracy for longer documents ~90%
- Lots of easy cases (… horrible… or … awesome …)

- For dataset of single sentence movie reviews (Pang and Lee, 2005) accuracy never reached above 80% for >7 years

- Harder cases require actual understanding of negation and its scope and other semantic effects

Data: Movie Reviews

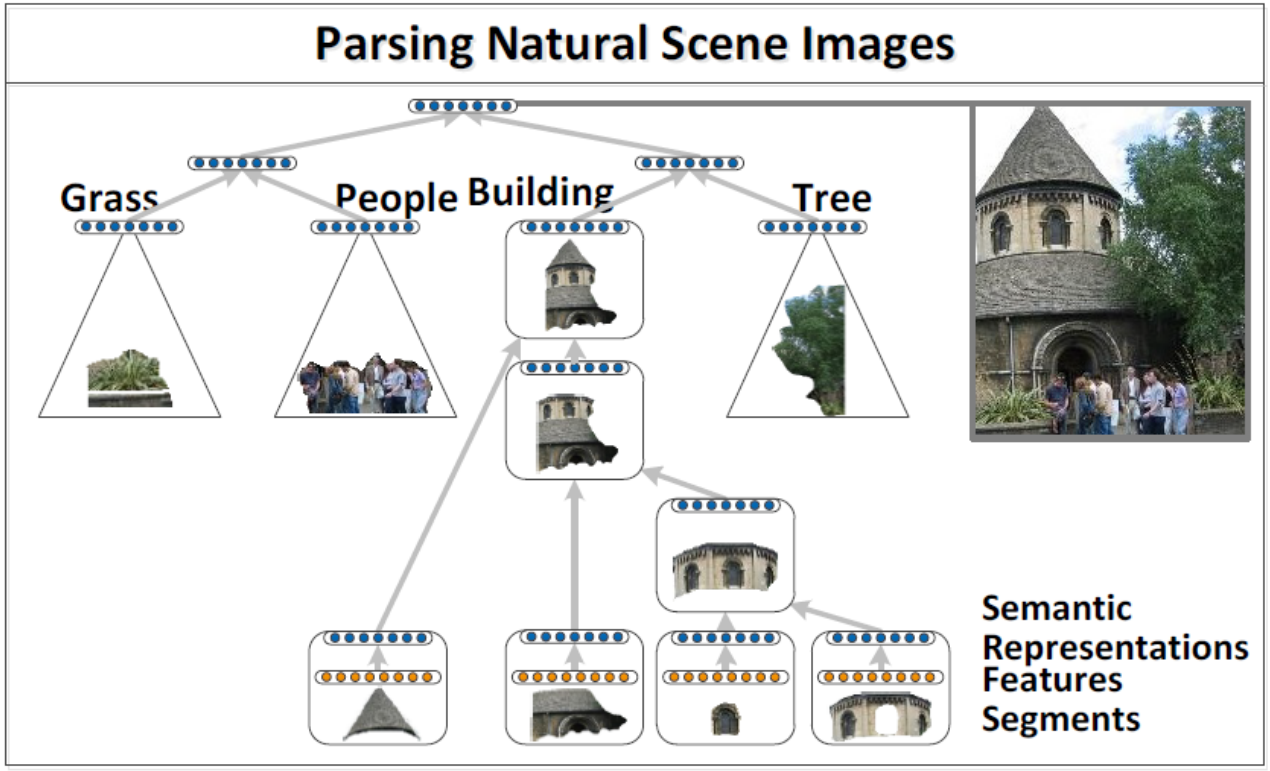
Stealing Harvard doesn't care about cleverness, wit or any other kind of intelligent humor.

There are slow and repetitive parts but it has just enough spice to keep it interesting.

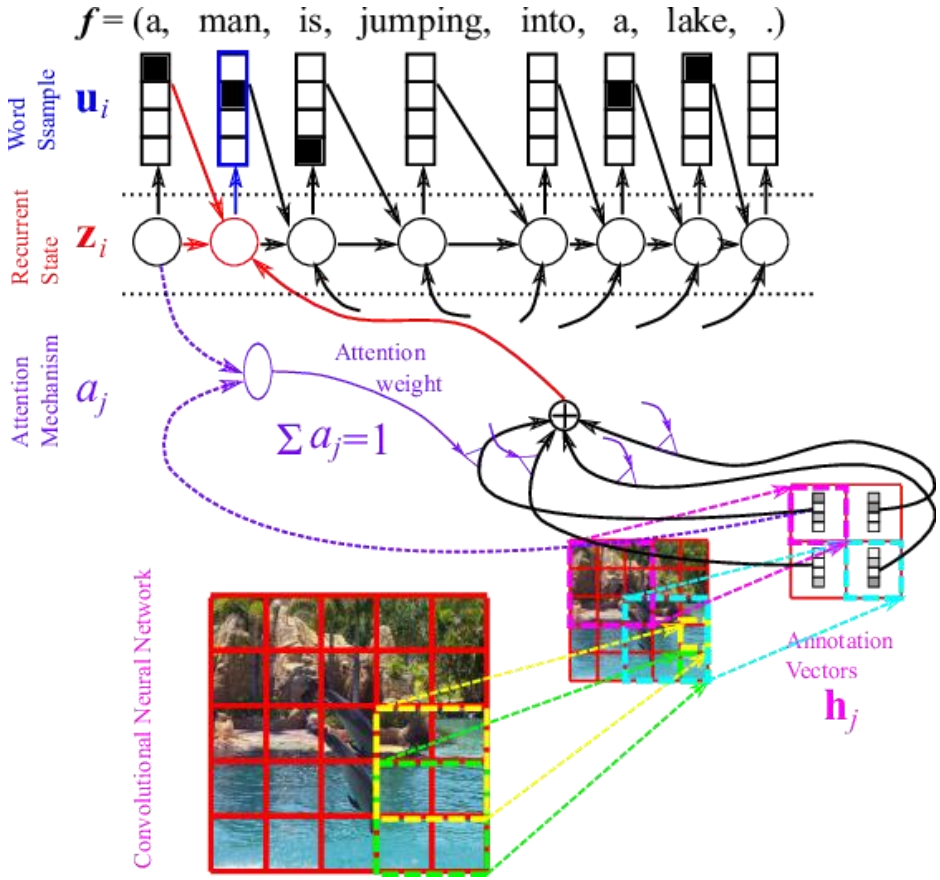
13 Linguagens naturais e imagens

Algorithm for Parsing Images

Same Recursive Neural Network as for natural language parsing!
(Socher et al. ICML 2011)



- Caption generation:



Fonte: https://devblogs.nvidia.com/introduction-neural-machine-translation-gpus-part-3/

Vinyals et al. (2015) “Show and Tell : A Neural Image Caption Generator, CVPR, http://static.googleusercontent.com/media/research.google.com/pt-BR//pubs/archive/43274.pdf

14 Modelo de linguagem de larga escala

- BERT (Bidirectional Encoder Representations from Transformers): modelo de linguagem de alto desempenho, lançado em outubro/2018 pelo Google, que representou um expressivo salto de qualidade no uso de aprendizado de máquina para NLP, cobrindo 103 línguas, pré-treinado com Wikipedia + BookCorpus e disponível em código aberto.
- Existem outros modelos de linguagem de larga escala, como GPT-2 e XL-Net, mas BERT representa o estado-da-arte em muitas soluções práticas de *natural language understanding* (NLU), ao alcance de indivíduos e corporações. No entanto, por se tratar de uma rede de grandes dimensões, há um desafio para viabilizar aplicações em tempo real, que requerem baixa latência.
- A evolução tecnológica tem contribuído para eliminar essa latência, por exemplo, com o lançamento pela NVIDIA de uma plataforma de alto desempenho

computacional para inferência em redes profundas, denominado TensorRT, que chega a ser 40 vezes mais rápida que o emprego apenas de CPUs.

- Isso tem viabilizado soluções de inteligência artificial para conversação em tempo real.
- BERT faz uso de um mecanismo de atenção chamado *Transformer* (VASWANI et al., 2017) que aprende relações de contexto entre palavras. O *Transformer* foi treinado com uma estrutura codificador-decodificador e apenas o codificador é necessário nas aplicações práticas, envolvendo outras tarefas de NLP.
- Diferente dos modelos direcionais, que trabalham com o texto de entrada de forma sequencial, o *Transformer* trabalha com a sequência inteira de uma vez. Por essa razão, ele é considerado bidirecional, mas não recorre à estratégia da Seção 9, devotada a modelos direcionais.
- Com essa estratégia, o contexto de uma palavra é capturado levando-se em conta todo o seu entorno.

Context is Everything

- No Context (Word2Vec)
  - river [bank]
  - [bank] deposit
- Left-to-Right Context (RNN)
  - I made a [bank] deposit
  - I made a [...]
- Bidirectional Context (?)
  - I made a [bank] deposit
  - I made a [...] deposit

Fonte: https://www.youtube.com/watch?v=BhlOGGzC0Q0

- BERT_base tem 12 camadas e 110 milhões de pesos sinápticos, enquanto BERT_large apresenta 24 camadas e 345 milhões de pesos sinápticos.
- Vimos que muitos modelos de linguagem aprendem a predizer a próxima palavra numa sequência. Trata-se de uma abordagem direcional que restringe / polariza o aprendizado de contexto. Visando ganho de desempenho, BERT adota duas estratégias de aprendizado:
  - ✓ *Masked Language Model* (MLM): Antes de alimentar o BERT com sentenças, 15% das palavras são substituídas por uma máscara (na prática, é um pouco mais elaborado que isso). O modelo, então, busca predizer o valor original das palavras que foram mascaradas, tomando por base o contexto produzido pelas palavras não-mascaradas. Na verdade, todas as palavras são preditas, mas a função de perda apenas considera o desempenho voltado para as palavras mascaradas.

- ✓ *Next Sentence Prediction* (NSP): o BERT é alimentado por pares de sentenças e cabe ao modelo predizer se a segunda sentença do par é de fato uma sentença subsequente no documento original. Metade das sequências alimentadas são pares subsequentes e a outra metade não.

- Para mais detalhes, consultar DEVLIN et al. (2019).

15 Referências bibliográficas

BAHDANAU, D.; CHO, K.; BENGIO, Y. "Neural Machine Translation by Jointly Learning to Align and Translate", 3rd International Conference on Learning Representations (ICLR), 2015.

BARONI, M.; JOULIN, A.; JABRI, A.; KRUSZEWSKI, G.; LAZARIDOU, A.; SIMONIC, K.; MIKOLOV, T. "CommAI: Evaluating the first steps towards a useful general AI", arXiv:1701.08954, 2017.

BENGIO, Y.; DUCHARME, R.; VINCENT, P. "A neural probabilistic language model", Journal of Machine Learning Research, vol. 3, pp. 1137-1155, 2003.

CHO, K.; VAN MERRIENBOER, B.; GULCEHRE, C.; BAHDANAU, D.; BOUGARES, F.; SCHWENK, H.; BENGIO, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724-1734, 2014 (também disponível em [https://arxiv.org/abs/1406.1078]).

DEERWESTER, S.; DUMAIS, S.T.; FURNAS, G.W.; LANDAUER, T.K.; HARSHMAN, R. "Indexing by latent semantic analysis", Journal of the Association for Information Science and Technology, vol. 41, no. 6, pp. 391-407, 1990.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", arXiv:1810.04805v2, May 2019.

GÉRON, A. "Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow", O'Reilly, 2019.

IRSOY, O.; CARDIE, C. "Deep recursive neural networks for compositionality in language", Neural Information Processing Systems (NIPS), 2014.

LANDAUER, T.K.; FOLTZ, P.W.; LAHAM, D. "An introduction to latent semantic analysis", Discourse Processes, vol. 25, no. 2-3, pp. 259–284, 1998.

LECUN, Y.; BENGIO, Y.; HINTON, G. "Deep learning", Nature, vol. 521, pp. 436-444, 2015.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. "Efficient Estimation of Word Representations in Vector Space", In ICLR Workshop Papers, 2013a (também disponível em [https://arxiv.org/abs/1301.3781]).

MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G.; DEAN, J. "Distributed representations of words and phrases and their compositionality", Neural Information Processing Systems (NIPS), pp. 3111-3119, 2013b.

PAPINENI, K.; ROUKOS, S.; WARD, T.; ZHU, W.J. "BLEU: a method for automatic evaluation of machine translation", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311-318, 2002.

PAULUS, R.; XIONG, C.; SOCHER, R. "A Deep Reinforced Model for Abstractive Summarization", arXiv:1705.04304, 2017.

SOCHER, R.; MANNING, C.D. "Deep learning for NLP (without magic)", Tutorial at The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, [https://nlp.stanford.edu/courses/NAACL2013/], 2013.

SUTSKEVER, I.; VINYALS, O.; LE, Q.V. "Sequence to Sequence Learning with Neural Networks", Advances in Neural Information Processing Systems 27 (NIPS), 2014 (também disponível em [https://arxiv.org/abs/1409.3215]).

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A.N.; KAISER, L.; POLOSUKHIN, I. "Attention Is All You Need", arXiv:1706.03762v5, 2017.