Strojový překlad pomocí neuronových sítí

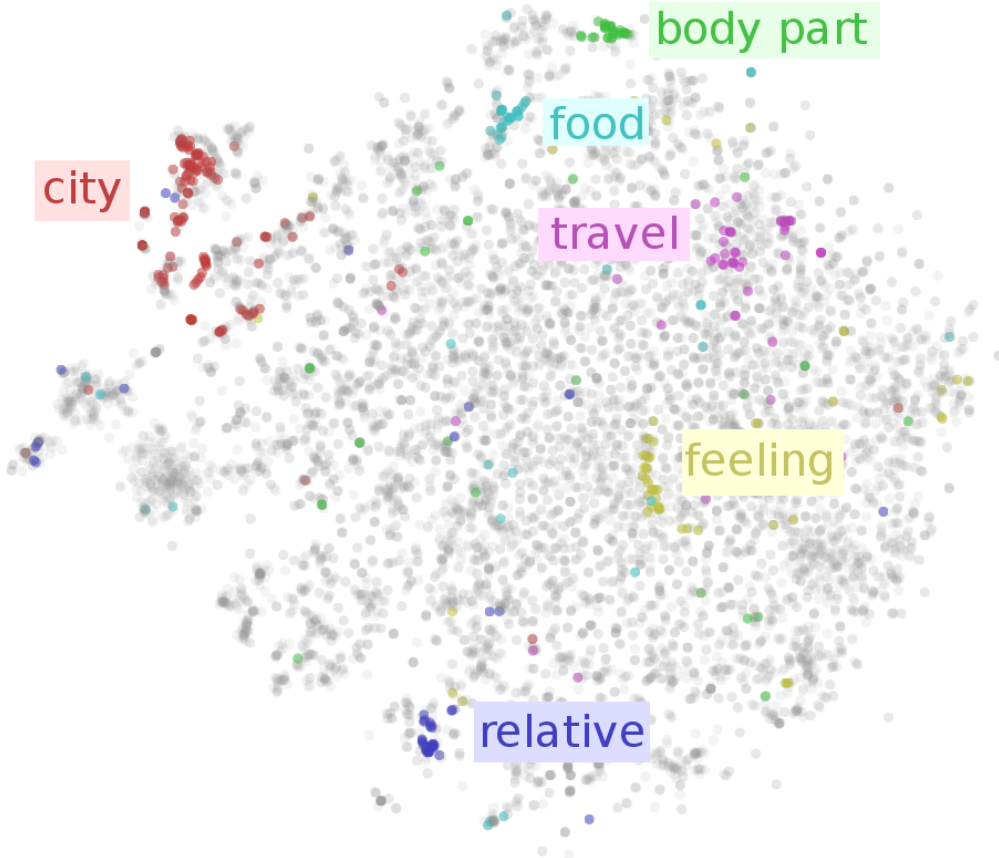
Martin Wörgötter

Úloha strojového učení

- Učení s učitelem
- Zdroje dat: literatura, Wikipedie, legislativa EU, dokumentace k počítačovým nástrojům, vícejazyčné webové stránky, syntetická data ...
- Soubor dvojic vět (zdroj, cíl)
- Segmentace na části (slova, podslova, znaky)
- Omezená velikost slovníku
- Chybová funkce $-log(p)$
- Výpočet nejpravděpodobnější cílové sekvence na základě zdroje

Kódování vstupu

- Každé slovo ve větě reprezentuje *vložení*



- Vektor $\mathbb{R}^n$, $n \in \mathbb{N}$

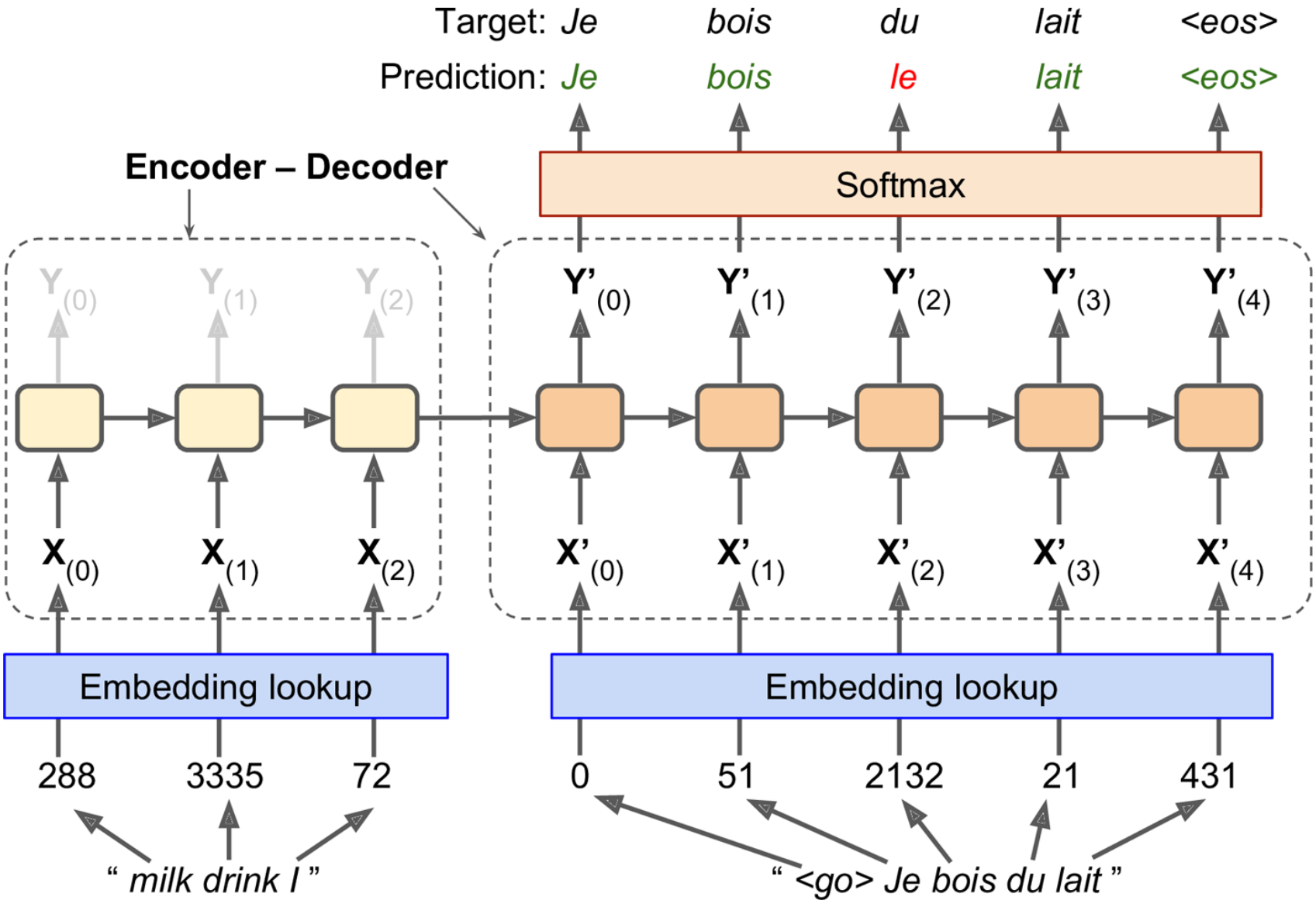
Zdroj: http://colah.github.io/posts/2015-01-Visualizing-Representations/

Rekurentní neuronová síť

- Kodér-dekodér
- Vstupy $x_t = (x_{t1}, \ldots, x_{tN})$ pro $t = 1, \ldots, M$
- Stav sítě

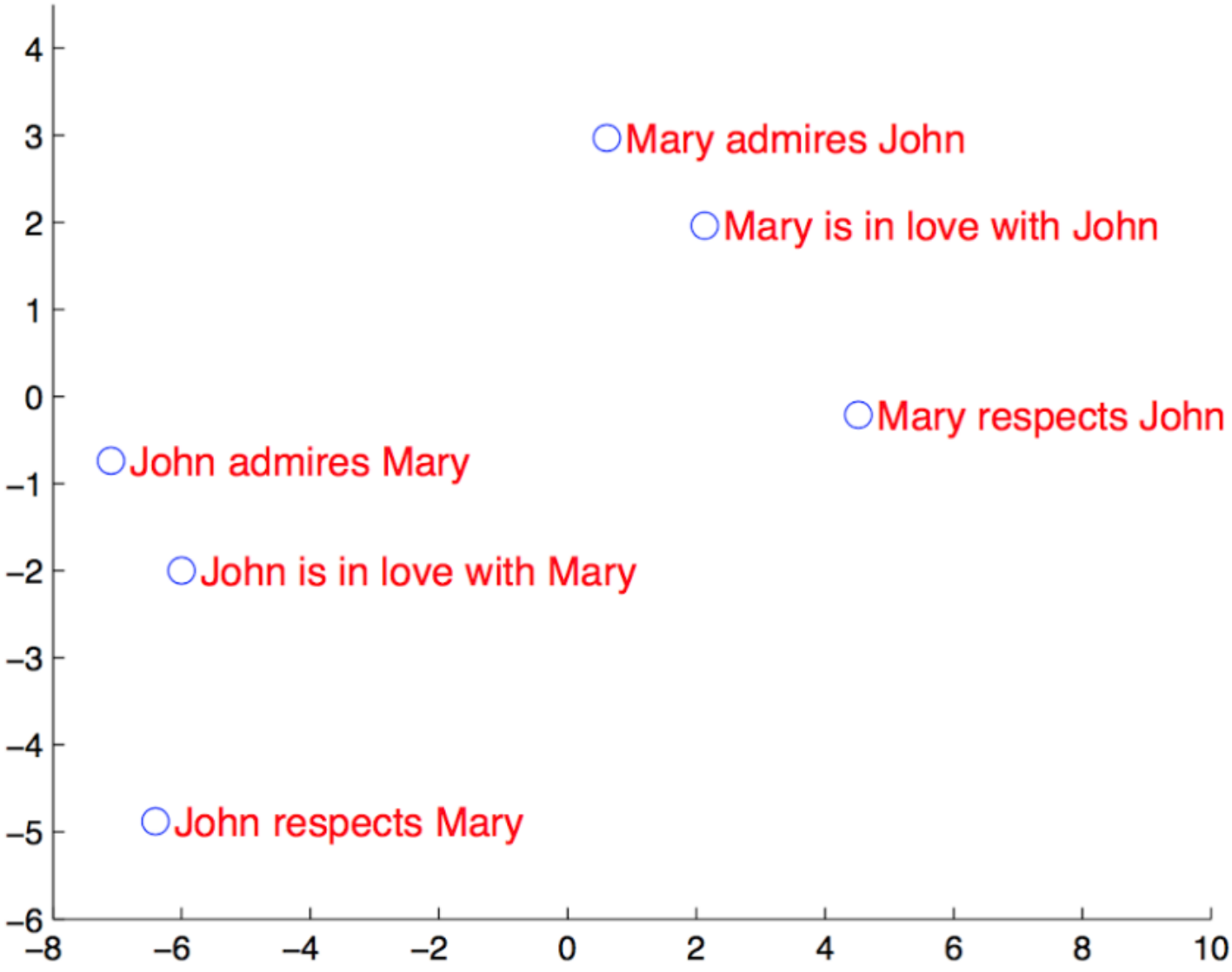
$$h_t = f(UCx_t^{\mathsf{T}}, Wh_{t-1})$$

- Výsledný stav kodéru je vstupem pro dekodér
- Analogicky probíhá výpočet dekodéru, stav je podmíněn vlastním výstupem
- Nadstavba dekodéru (softmax) převádí stav na pravděpodobnostní rozdělení nad slovníkem cílového jazyka
- Učení algoritmem *SGD*



Zdroj: https://www.safaribooksonline.com/library/view/neural-networks-and/9781492037354/ch04.html

Projekce kontextového vektoru do 2D



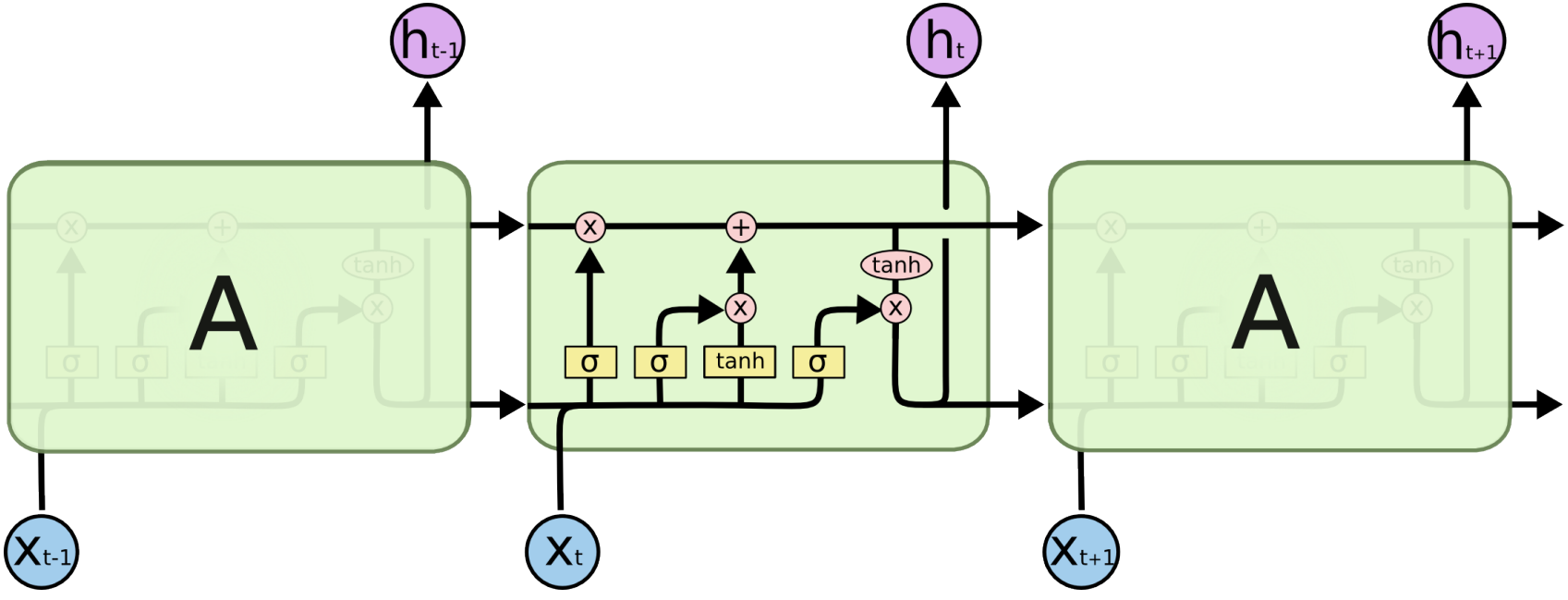
Zdroj: https://devblogs.nvidia.com/introduction-neural-machine-translation-gpus-part-2/

Předzpracování dat

- Tokenizace
- BPE: Gesundheits|forsch|ungsinstitu|ten
- Normalizace
- Značkování: morfologické, syntaktické, ...
- Odstranění šumu
- Zdrojový text v opačném pořadí

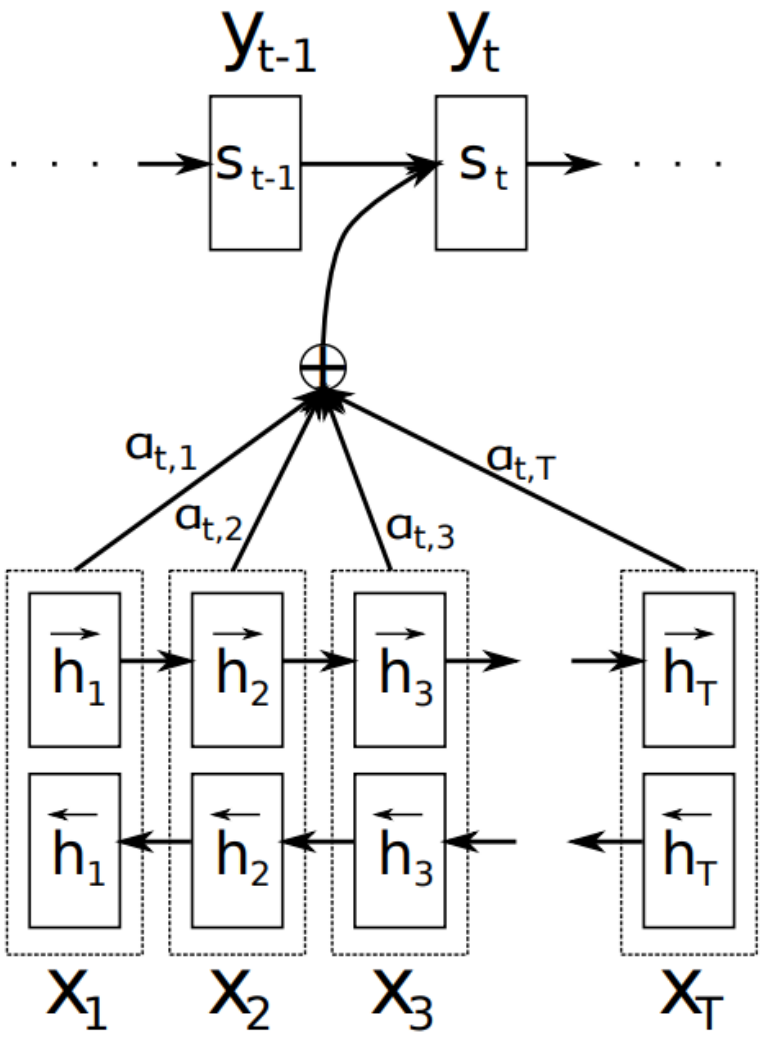
Rozšíření modelu

Long Short Term Memory (LSTM)



Zdroj: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

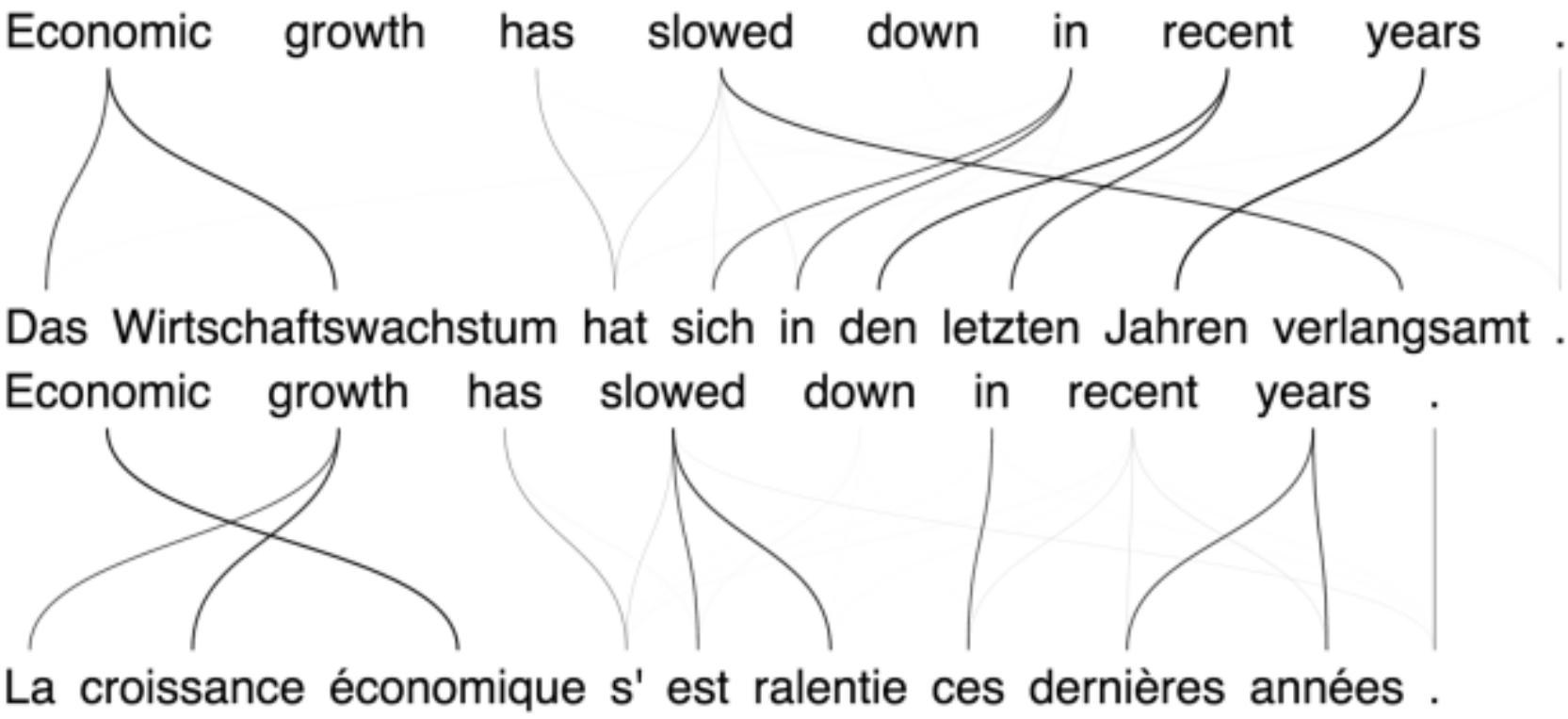
Zarovnávací model (Bahdanau et alii 2015)



- Obousměrná rekurentní neuronová síť
- Vstup je čten v obou směrech a stavy odpovídají spojení vektorů
- Pravděpodobnostní rozdělení nad množinou stavů kodéru
- Vážený součet stavů kodéru

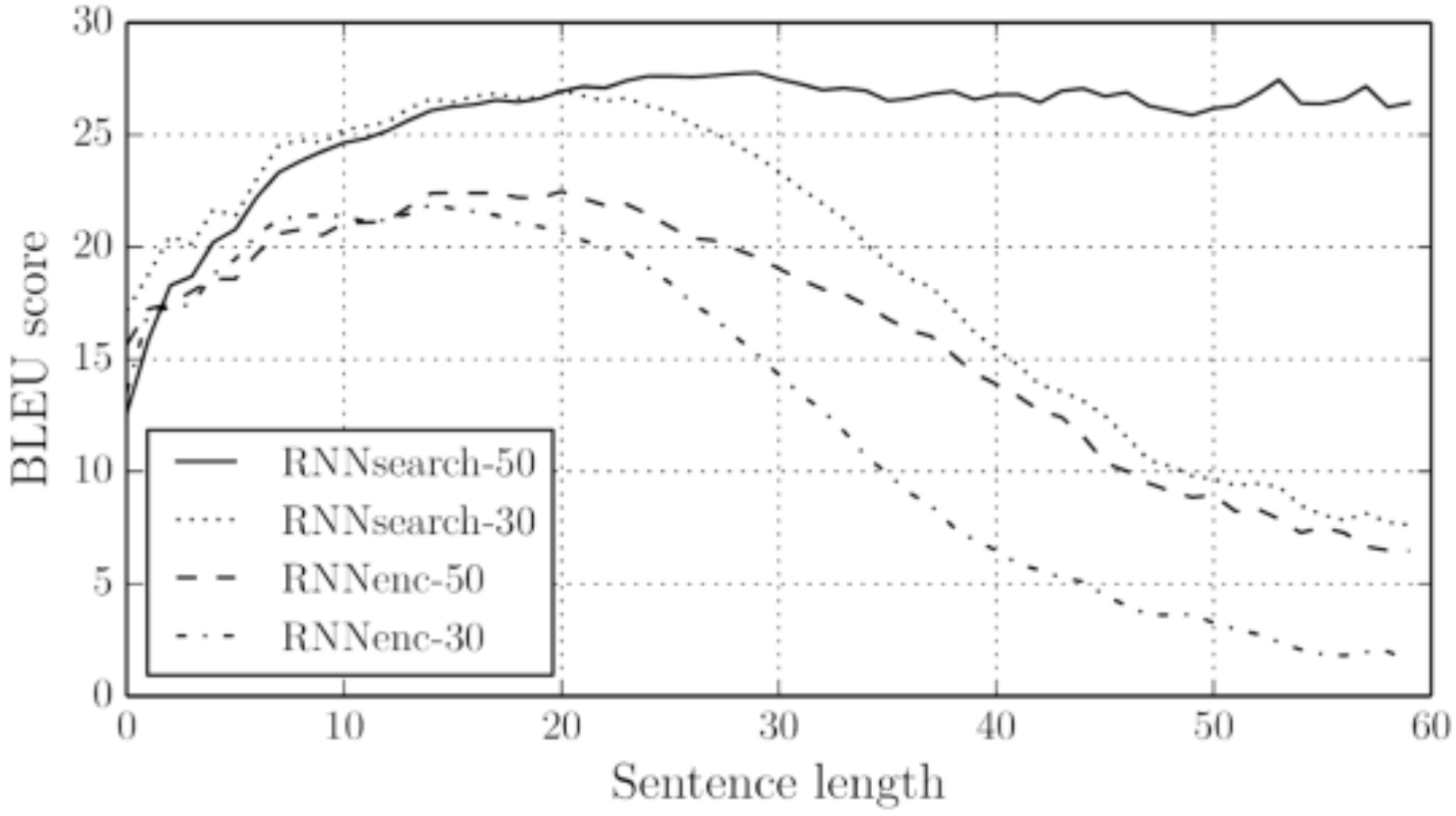
Zdroj:
https://machinelearningmastery.com/global-attention-for-encoder-decoder-recurrent-neural-networks/

Zarovnání na překladu NJ $\rightarrow$ AJ, FRJ $\rightarrow$ AJ



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

Kvalita překladu v závislosti na délce vět



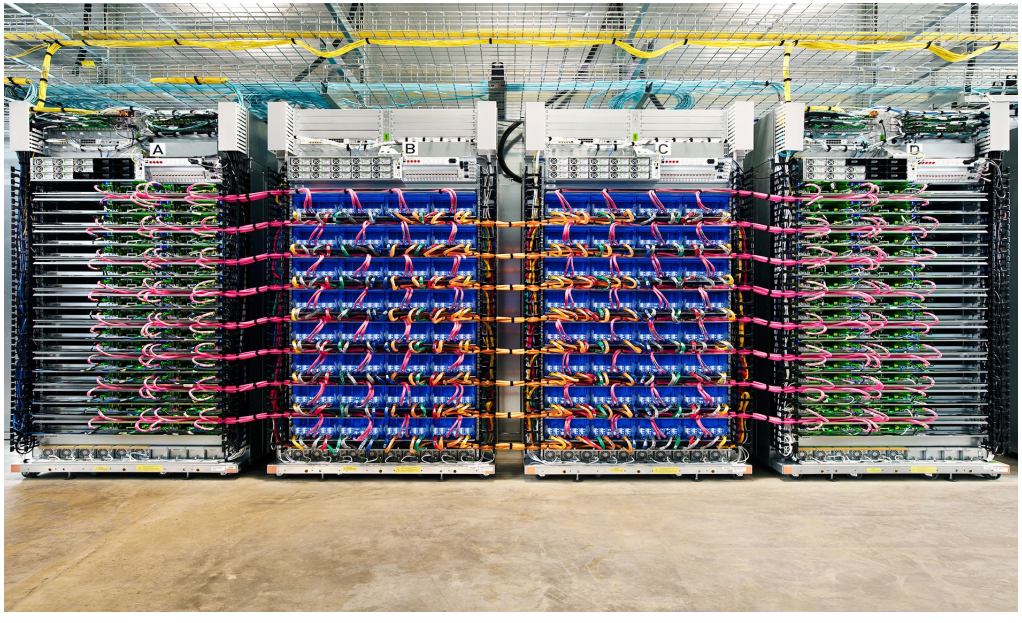
Zdroj: https://devblogs.nvidia.com/introduction-neural-machine-translation-gpus-part-3/

Učení hlubokých neuronových sítí

- Učení rekurentní neuronové sítě nelze jednoduše paralelizovat
- Paralelizace skrze učení více kopií neuronové sítě
- Grafické karty podporují efektivní práci s vektory a maticemi
- Dávkové zpracování
- Na výkonných grafických kartách běží výpočty několik dní
- Automatická evaluace, metrika *BLEU*

Google Tensor Processing Unit (TPU)

- 8-bitová kvantizace
- Efektivní maticové násobení
- CISC instrukční sada
- Výkon 180 Teraflopů a 64 GB paměti
- Konfigurace 64 $\times$ TPU v2



Zdroj: http://www.starkinsider.com/2017/05/google-placing-big-bets-ai-machine-learning.html

Ukázky strojového překladu – Google Translate

Zdroj: Engine and boiler rooms and other premises in which flammable or toxic gases are likely to escape shall be capable of being adequately ventilated.

Překlad: K motorovým a kotelenským místům a jiným prostorům, kde by pravděpodobně unikly hořlavé nebo toxické plyny, musí být dostatečně větrané.

Referenční překlad: Strojovny a kotelny a další prostory, ve kterých se mohou uvolňovat zápalné nebo jedovaté plyny, musí být přiměřeně odvětrávány.

Ukázky strojového překladu – Google Translate

Zdroj: The Agency is managed by its Executive Director, who is independent in the performance of his duties.

Překlad: Agenturu řídí výkonný ředitel, který je při plnění svých povinností nezávislý.

Referenční překlad: Agenturu řídí výkonný ředitel, který je při výkonu svých povinností zcela nezávislý.

Rozdíl mezi obecným a doménově zaměřeným překladem

Zdroj: Silk waste (including cocoons unsuitable for reeling), yarn waste and garneted stock, other than not carded or combed

Překlad (Google Translate): Hedvábný odpad (včetně kokosů nevhodných pro navíjení), odpad z příze a jiné než ne mykané nebo nečesané

Překlad (doménově specifický): hedvábný odpad (včetně zámotků nevhodných ke smotávání), přízový odpad a rozvlákněný materiál, jiný než nemykaný a nečesaný
= **referenční překlad**

Děkuji za pozornost