Semestrální projekt

Strojový překlad pomocí umělých neuronových sítí

Jonáš Holcner

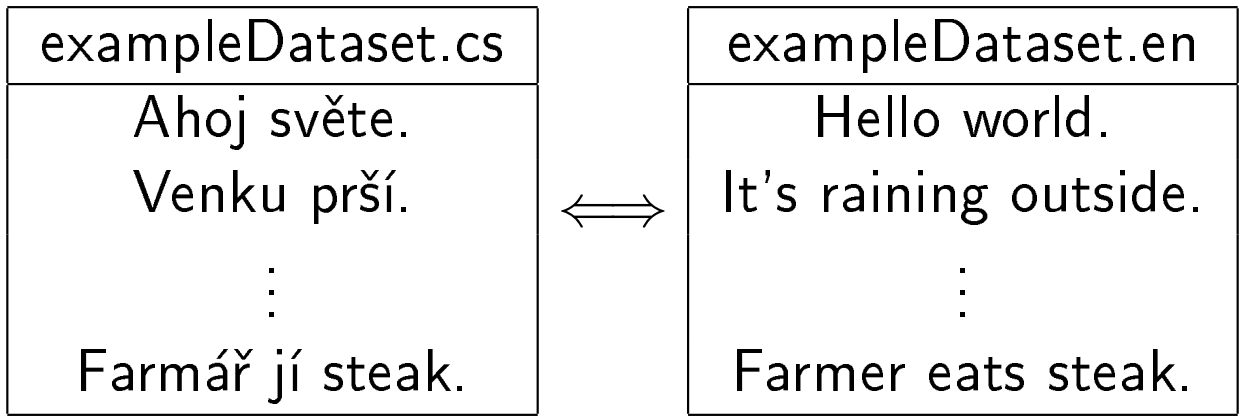
Vedoucí: Ing. Igor Szőke, Ph.D.

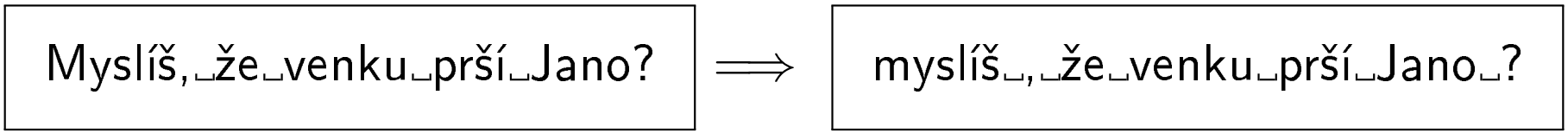
23. ledna. 2018

Cíle na zimní semestr

1. Nastudovat teorii a zvolit vhodný nástroj pro vývoj překladače
2. Najít, zvolit a připravit vhodná data pro trénování
3. Vytvořit, natrénovat a otestovat překladač pro překlad z jednoho do druhého jazyka

Datasety

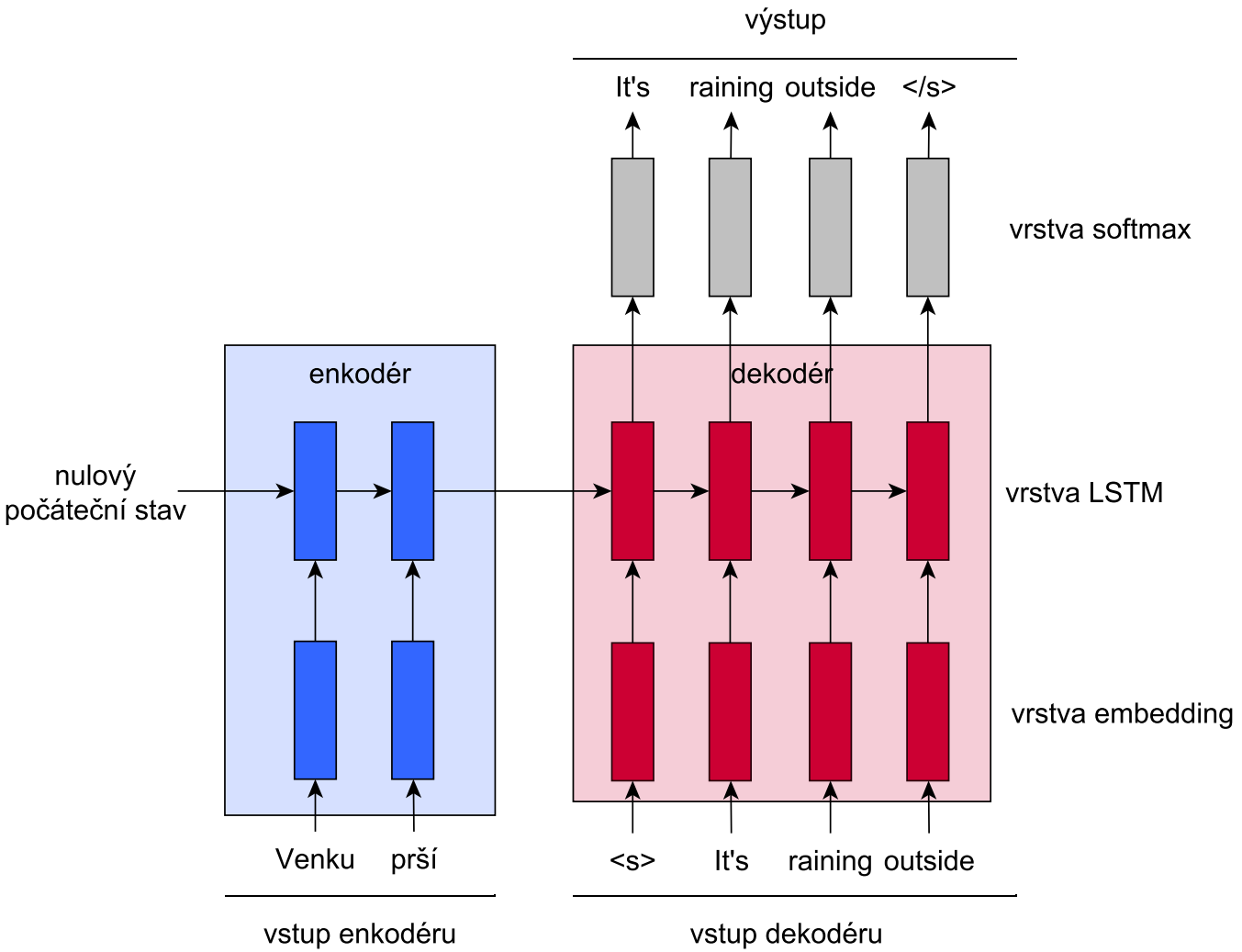




Obrázek: Předzpracování - tokenizace a truecasing

Strojový překlad pomocí neuronových sítí

- Překlad po celých sekvencích (seq2seq)
- Rekurentní neuronové sítě (LSTM)
- Word embeddings



Co jsem udělal

- Dataset WMT newsCommentary2012 pro trénink, WMT newsTest2017 pro testování. Předzpracování pomocí skriptů z Moses
- CS $\Rightarrow$ EN
- Baseline systém v nástroji pro statistický překlad Moses
- Překladový systém (balíček *nmt* v Python + Keras)
- bit.do/pythonNmt

Výsledky

	baseline v Moses	vytvořený systém
BLEU skóre	14.0	2.03

Obrázek: Porovnání pro dataset WMT newtest2017 CS $\Rightarrow$ EN

osmadvacetiletý šéfkuchař nalezen mrtev v obchodě v San Francisku

$\Downarrow$

_UNK _UNK _UNK _UNK in _UNK in China

Obrázek: Ukázka překladu z testovacího datasetu WMT newtest2017

Plán na letní semestr

- Natrénovat síť přes několik jazyků najednou
- Při tokenizaci místo celých slov použít části slov (BPE)
- Použít větší datasety (titulky)
- Přidat více vrstev RNN a obousměrný enkodér
- Použit jinou techniku pro výběr slov při generování (beam search místo greedy search)