Neuronové sítě a strojový překlad

Jindřich Libovický, 3. listopadu 2022

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Obsah

1. Strojové učení
2. Neuronové sítě
3. Jak funguje strojový překlad
4. Ukrajinsko-český překladač

Ochutnávka toho, co se dá studovat v oborech umělá inteligence a jazykové technologie.

Umělá inteligence & strojové učení

Co je to umělá inteligence?

Navigace v mobilním telefonu

Rozpoznávání řeči:
Jak se dostanu odsud na Smíchov?

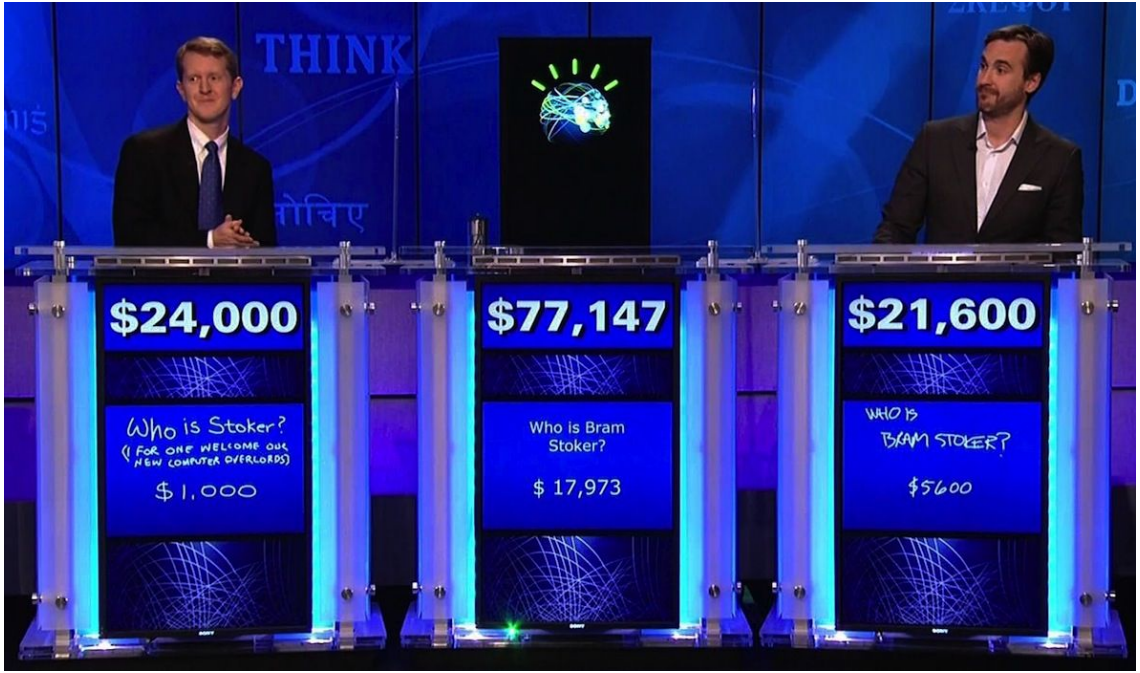
Gramatická pravidla:
intention=journey
destination=Smíchov

Vyhledání v databázi
a optimalizace cesty

Kasparov vs. Deep Blue (1997)

… žádná inteligence, jen hrubá výpočetní síla.

Watson (2011)



… žádná inteligence, jen chytře indexovaný text.

Lee Se-dol vs. AlphaGo (2016)

Umělá inteligence!

Lee Se-dol vs. AlphaGo (2016)

Technet.cz

iDNES.cz | Zprávy | Kraje | Sport | Kultura | Ekonomika | Bydlení | Technet | Ona | Revue | Auto |

Audio | Video | Tv | Foto | PC & Mac | Software | Notebooky | Web | Věda & Vesmír | Vojenství | Přelomo

Go má svého boha: umělá inteligence porazila nejlepšího hráče světa 3:0

25. května 2017 17:14, aktualizováno 30. května 11:12

Software AlphaGo si poradil s nejlepším hráčem go současnosti. V sérii na dva vítě
zápasy zvítězil celkem jednoznačně, jednotlivá střetnutí byla přesto zajímavá.

Kche Tie během druhého zápasu s AlphaGo | foto: Stringer , Reuters

V čínském Wu-čenu nedaleko Šanghaje se v druhé polovině května střetli dvě
největší hvězdy oblíbené východoasijské hry go: devatenáctiletý čínský
profesionální hráč Kche Tie (v anglickém přepisu Ke Jie) a software AlphaGo

Reklama

ČT24

SESTAVOVÁNÍ VLÁDY | DOMÁCÍ | SVĚT | REGIONY | EKONOMIKA | KU

Přelom: Umělá inteligence se za tři dny sama naučila go, pak porazila nejlepšího hráče světa

19. 10. 2017

Vloni v březnu porazil program AlphaGo nejlepšího světového hráče ve hře go. Zaskočilo to tehdy všechny, jak hráče Go, tak programátory – ukázalo to, jak rychlý je pokrok v tomto oboru a jak rychle stroje člověka dohánějí a nyní už i překonávají. Nyní vědci přišli s novou verzí programu AlphaGo, která tu starší porazila 100:0. A to přesto, že hru go znala pouhé tři dny.

V čem je rozdíl?

AlphaGo je neuronová síť
a učí se sama z dostupných dat.

Učení vs. programování řešení

Programování řešení

- Jsme schopni problém **formálně popsat** jasnými koncept
- Program je jednoznačný **návod**, jak s koncepty zacházet

Příklad - E-shop: *Koncepty: zboží, sklad, zákazní, objednávka*

Udělat objednávku, odeslat objednávku = jednoduchý algoritmus

Učení řešení

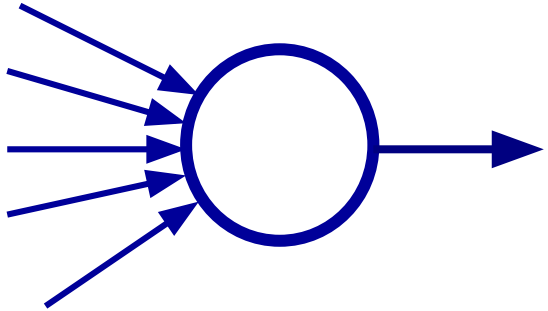
- Máme **příklady** vstupů a výstupů a **metriku**, jak dobré je řešení
- Nejsme schopni do důsledku napsat návod, jak úlohu řešit

Příklad - překlad: *Neexistuje žádný jednoduchý návod*

Existuje mnoho přeložených textů, co se dají použít k trénování

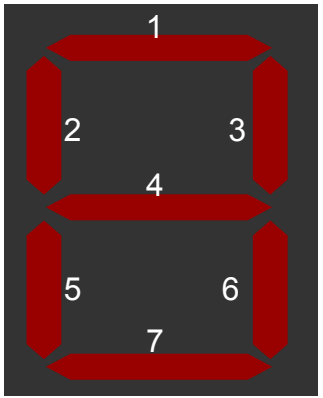
Neuronové sítě

Perceptron: nejjednodušší neuronová síť



- Velmi jednuduchý (a špatný) model neuronu
- Když je vážený součet vstupů větší než 1,
  vydá 1, jinak vydá 0
- Váhy ve váženém součtu se učí podle příkladů

Příklad: Detekce čísla 7 (1)

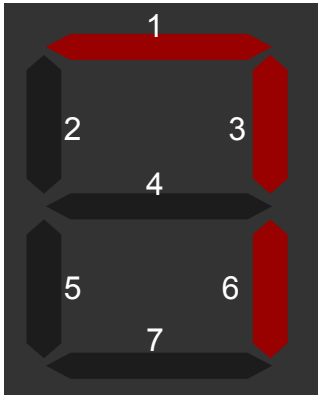


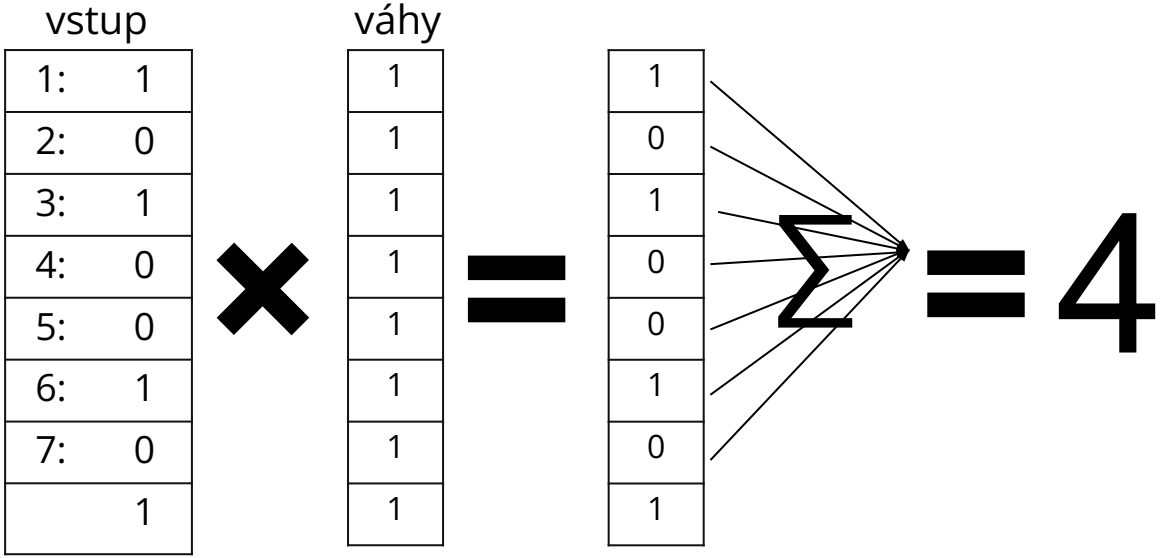
vstup		váhy
1:	?	1
2:	?	1
3:	?	1
4:	?	1
5:	?	1
6:	?	1
7:	?	1

× Σ = ?

Detekce čísla 7: když je výstup > 0
Perceptronový algoritmus,
nejjednodušší způsob trénování neuronové sítě

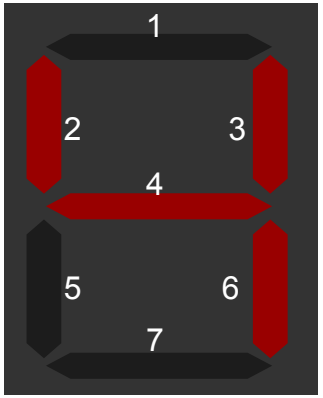
Příklad: Detekce čísla 7 (2)

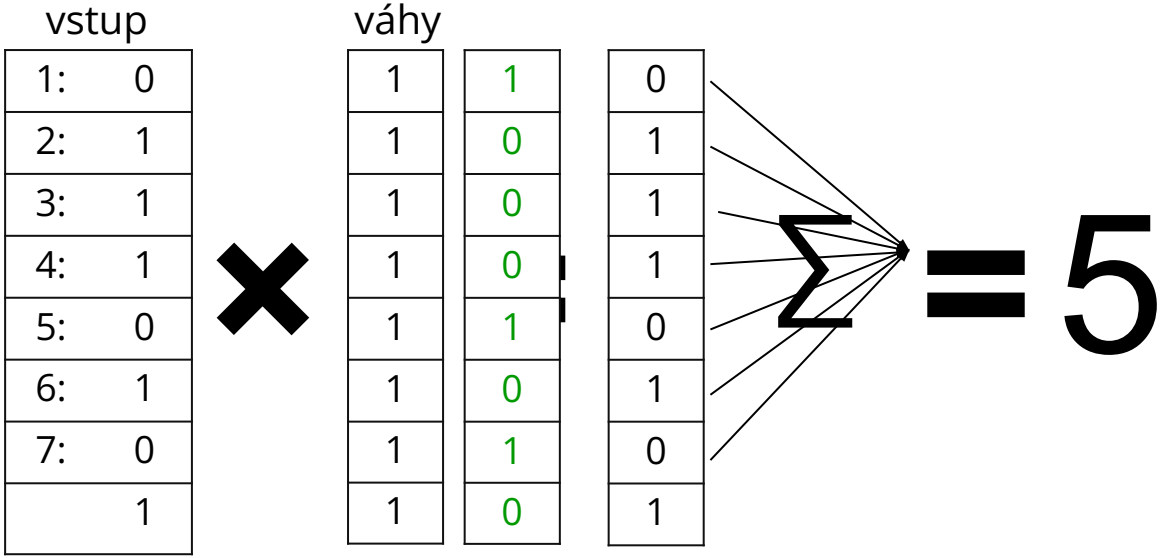




Na začátku všechny váhy 1
Zkusíme 7 na vstupu,
4 > 0, výsledek je správně

Příklad: Detekce čísla 7 (3)

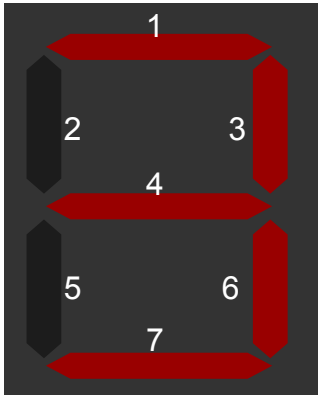


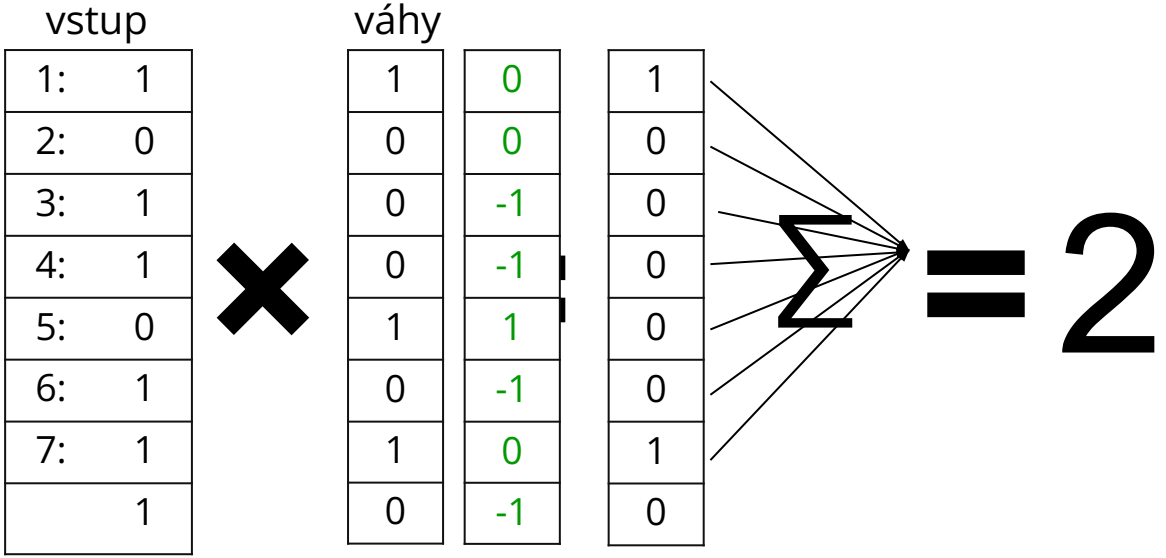


5 > 0, špatně

Upravíme váhy tak, že odečteme vstup

Příklad: Detekce čísla 7 (4)



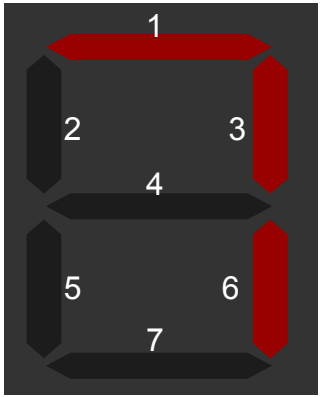


2 > 0, špatně

Upravíme váhy tak, že odečteme vstup

Konečně se objevují záporná čísla

Příklad: Detekce čísla 7 (5)



vstup	
1:	1
2:	0
3:	1
4:	0
5:	0
6:	1
7:	0
	1

×

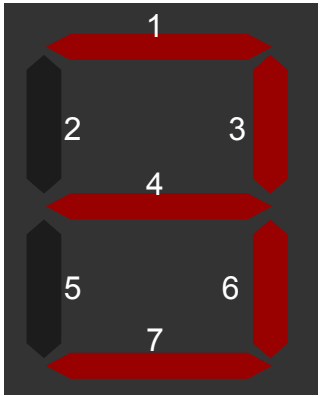
váhy		
0	1	1
0	0	0
-1	0	-1
-1	-1	0
1	1	0
-1	0	-1
0	0	0
-1	0	-1

$$\sum = -2$$

-2 < 0, špatně

Upravíme váhy tak, že **přičteme** vstup

Příklad: Detekce čísla 7 (6)

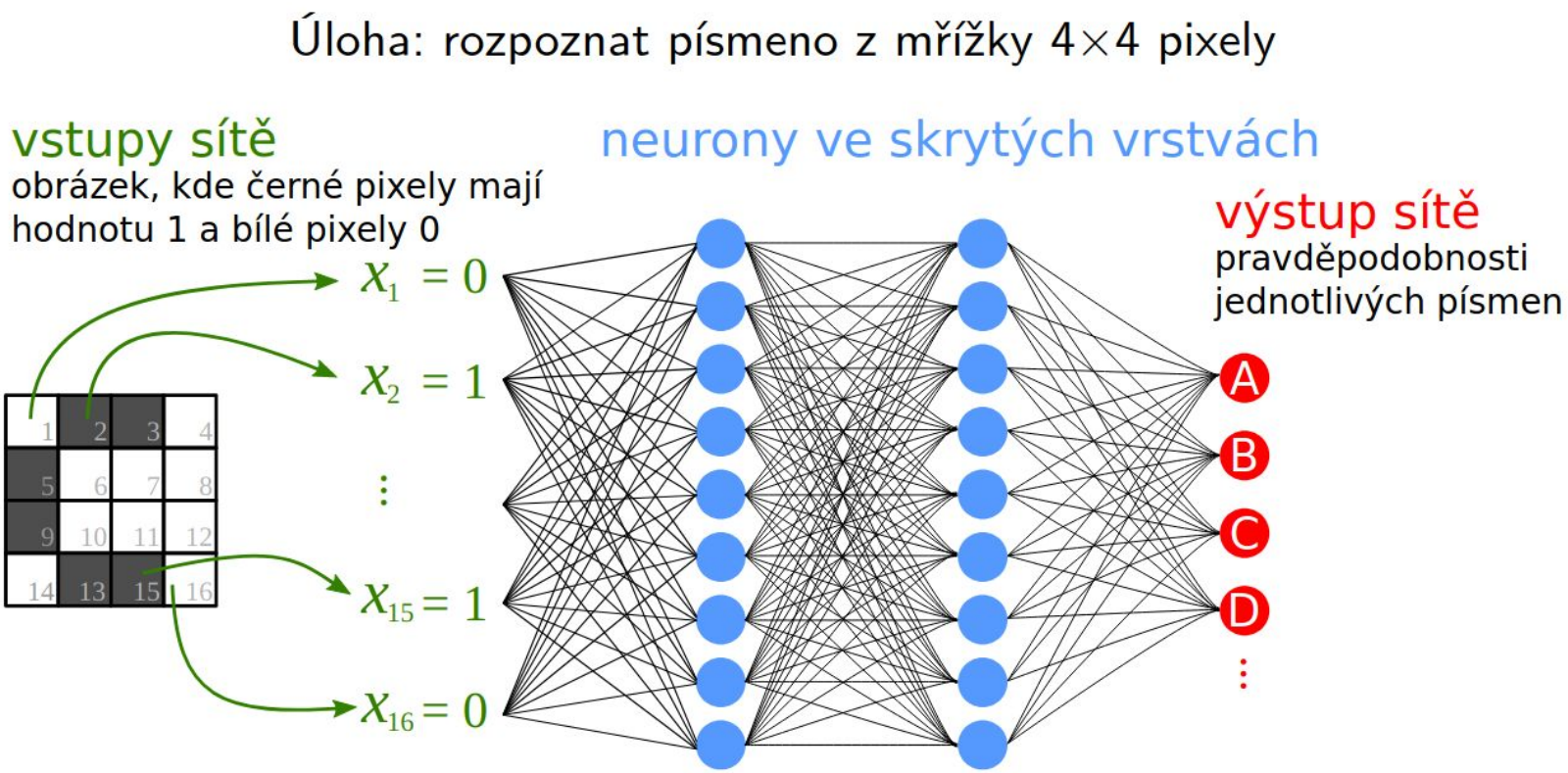


vstup		váhy			
1:	1	1	×	1	=
2:	0	0		0	
3:	1	0		0	
4:	1	-1		0	
5:	0	1		0	
6:	1	0		-1	
7:	1	0		0	
	1	0		0	

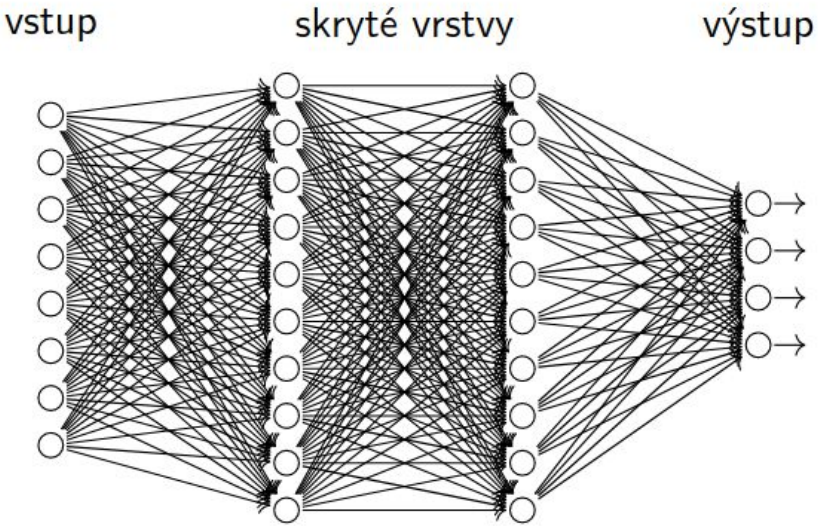
$\sum = 0$

0, správně

Neuronová síť: více vrstev



Neuronová síť: trocha matematiky



- Organizace do **vrstev**
  - Čísla na vstupu vrsty = vektor
  - Vážení vstupů = maticové násobení
  - Výstup vrstvy = vektor
- Většina výpočetů maticové násobení => rychlé počítání na **grafických kartách**
- Výstup NN = **spojitá funkce** vstupů

Učení neuronové sítě (poslední trocha matematiky)

- Trénování = **minimalizace chyby** na trénovacích datech
- Když je chyba spojitá funkce, můžeme spočítat derivaci chyby vzhledem k parametrům
- Posunout parametry po směru derivace = snížit chybu

Strojový překlad

Strojový překlad

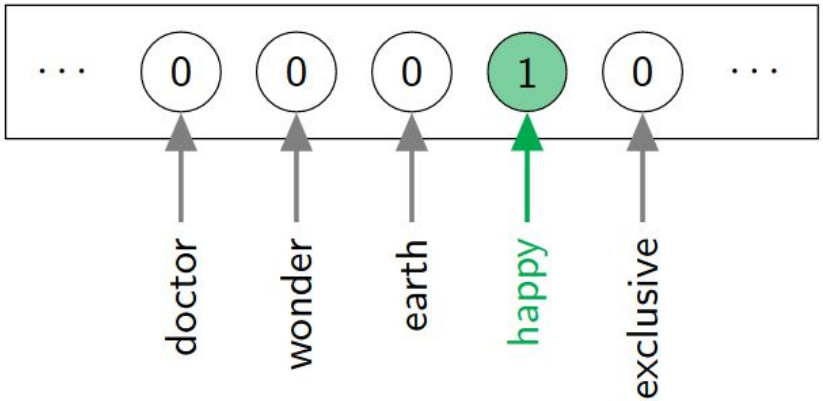
- 1950s - 1990s: snaha používat formální gramatiky a explicitní překladová pravidla

- 1990s - 2016: statistické modely založené na trénování z dat
  - Překladový model: tabulka, kde jsou možné překlady slov a frází
  - Jazykový model: říká, jak vypadá věta v cílovém jazyce
  - Dekódování: návrhy překladového modelu seřadit podle jazykového modelu

- Od 2016: neuronové sítě - obrovský skok v kvalitě
  - Poměrně složitá architektura...

Diskrétní slova do vstupních vektorů

- neuronové sítě potřebují spojité vstupy
- one-hot vektor: očíslujeme slova, 0/1 indikuje slovo

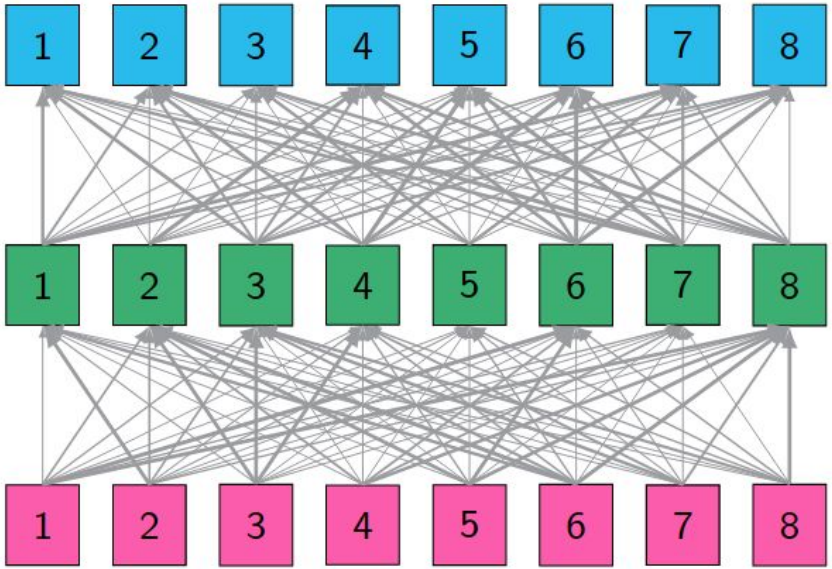


- první vrstva = násobení maticí

Každé **slovo** (nebo jiná jednotka) je reprezentovaná mnohodimenzionálním **vektorem**

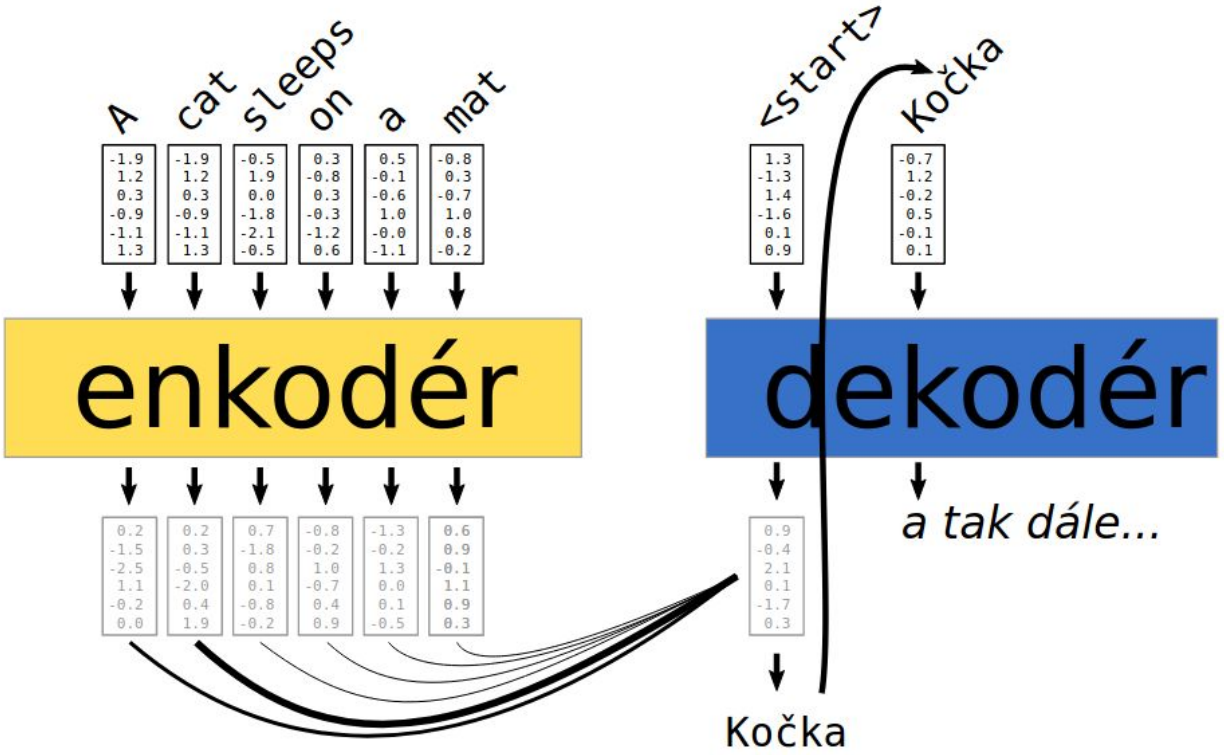
- učí se „mimochodem“ z dat
- mají zajímavé vlastnosti

Slova se musí zpracovat v kontextu



- Mezi klasickými vrstvami tzv. **self-attention**
- Každé slovo se „podívá" na ostatní slova a vezme si z něj relevantní informace
- Na začátku: vektor reprezentuje **izolované** slovo
  Na konci: vektor reprezentuje slovo **v kontextu** věty

Enkodér-dekodér a generování překladu



Dekodér sbírá informace **z předchozích slov** a **z enkodéru**

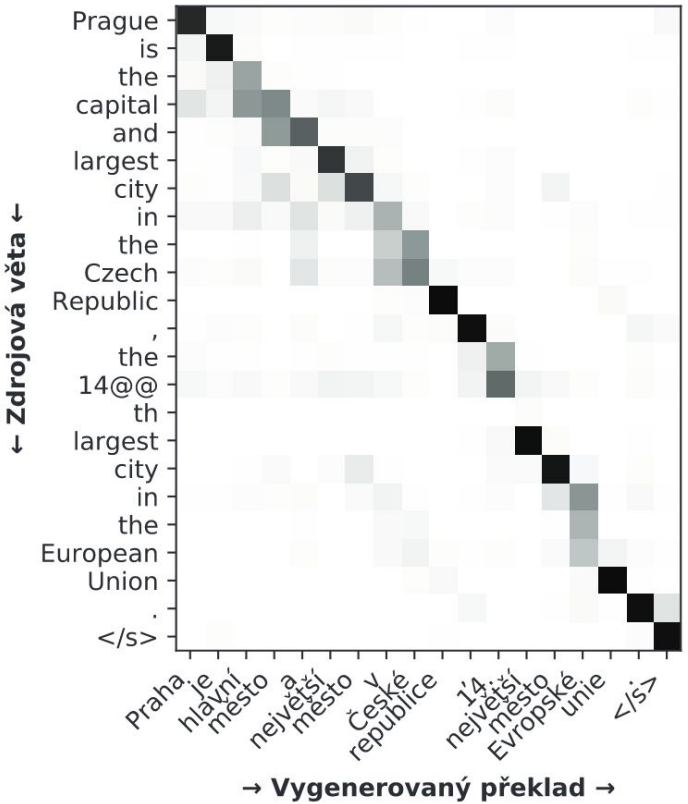
Kam se dívá dekodér, když generuje překlad

Modely se trénují **jenom na větách**

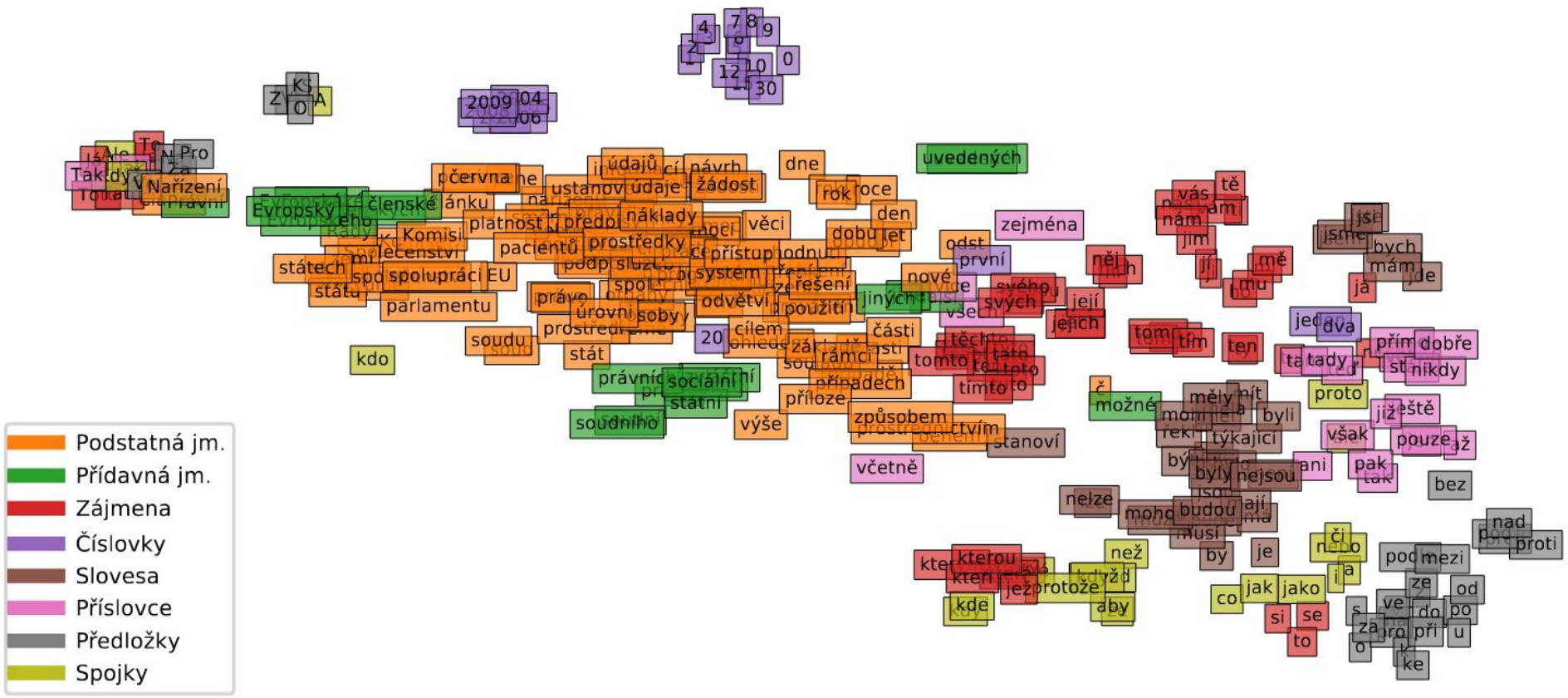
Na začátku neví **nic o slovech**

Z dat vyvodí:

1. Jak překládat věty
2. Jaké jsou vztahy mezi jednotlivými slovy v jednom jazyce
3. Jaké jsou vztahy mezi slovy ve zdrojovém a v cílovém jazyce



Naučené slovní reprezentace



Trénovací data

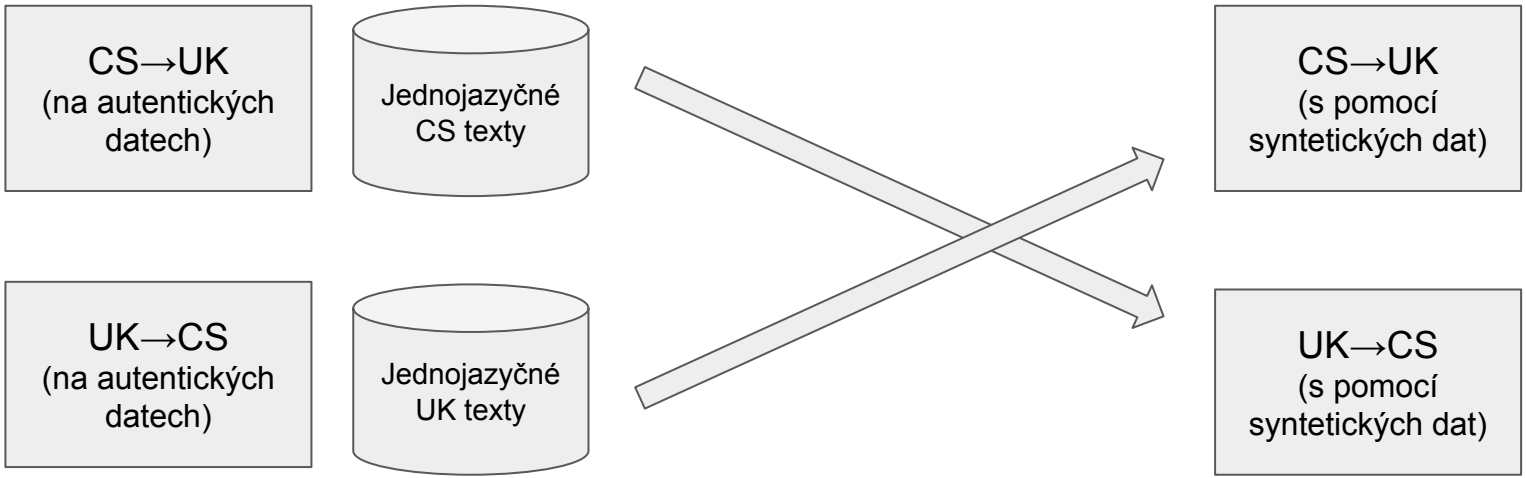
= Dvojice vět, které jsou navzájem svoje překlady

Nejčastější zdroje:

- Oficiální dokumenty EU, OSN…
- Filmové titulky
- Vícejazyčně weby
- Co nejvíce vět z webu a hledat mezi nimi potenciální překlady

Back-translation: možnost použít jednojazyčná data

- Analogie k AlphaGo: hraje hry samo se sebou
- Dva překladače si navzájem generují data



- Cílová strana je vždy autentický jazyk

Evaluace

- *Automatická* = **měření podobnosti** překladu s nějakým referenčním překladem
- *Ruční* = **lidé hodnotí** kvalitu překladu

WMT = Conference on Machine Translation

- Mezinárodní konference, kde probíhá mj. soutěž v kvalitě překladu
- Soutěží firmy i univerzity - většina systémů lepší než dostupné
- Díky MFF UK je vždy přítomná čeština, často vyhráváme

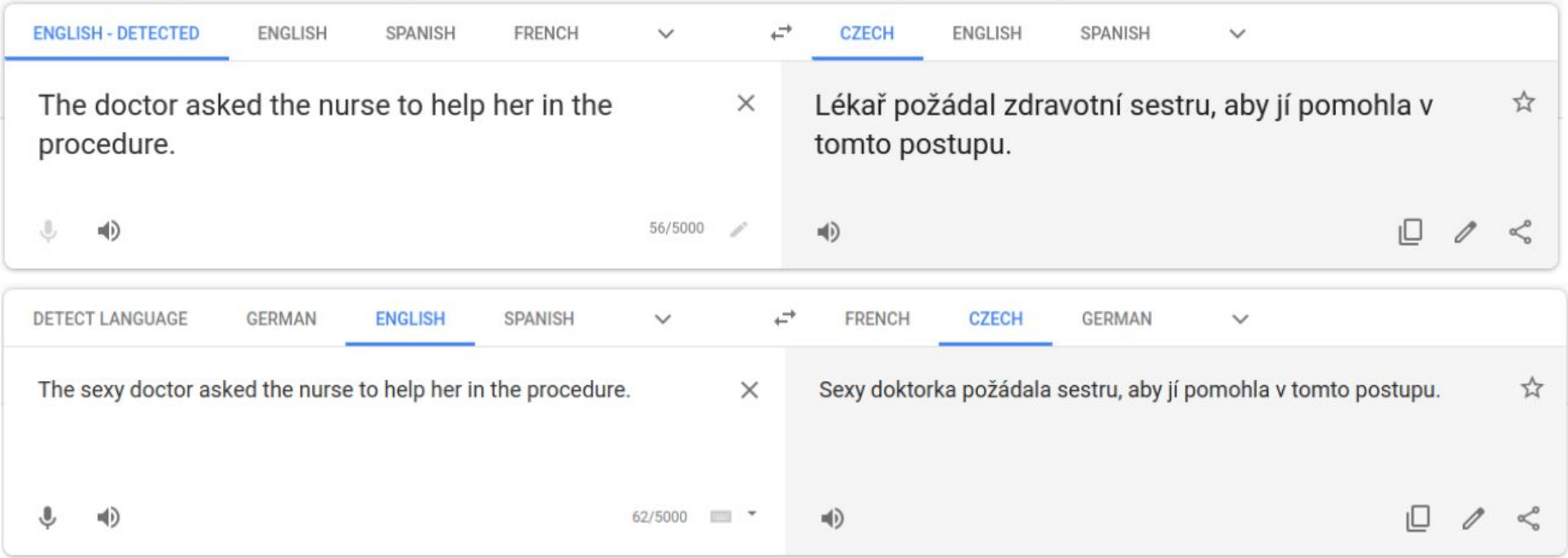
Často vyjde, že strojový překlad se kvalitou neliší od lidského.

Kde je problém?

Problémy strojového překladu

- Funguje dobře jenom pro jazyky, pro které máme hodně dat
  (evropské jazyky, čínština, japonština, arabština)
- Překládá se na úrovni vět, ne na úrovni dokumentů
- Problémy s terminologií a specifickou slovní zásobou
- Vyžaduje velký výpočetní výkon
- Zachycuje a zesiluje stereotypy v trénovacích datech

Stereotypy z dat se projevují i v modelech



Ukrajinsko-český překladač za tři týdny

Proč budovat vlastní systém (1)

- Google a Microsoft translate podporují cs-uk
- Mezistupeň angličtina = ztráta informací

- Funguje podobně
- Možná mezistupeň ruština?
- Nemusí být důvěryhodný pro uživatele

Proč budovat vlastní systém (2)

- WMT: soutěž v kvalitě strojového překladu
  Volně dostupné komerční systémy nejsou nejlepší
- ÚFAL dopadá v soutěžích dobře

nature communications

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature communications › articles › article

Article | Open Access | Published: 01 September 2020

Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals

Martin Popel ✉, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar & Zdeněk Žabokrtský

Nature Communications **11**, Article number: 4381 (2020) | Cite this article

27k Accesses | **28** Citations | **144** Altmetric | Metrics

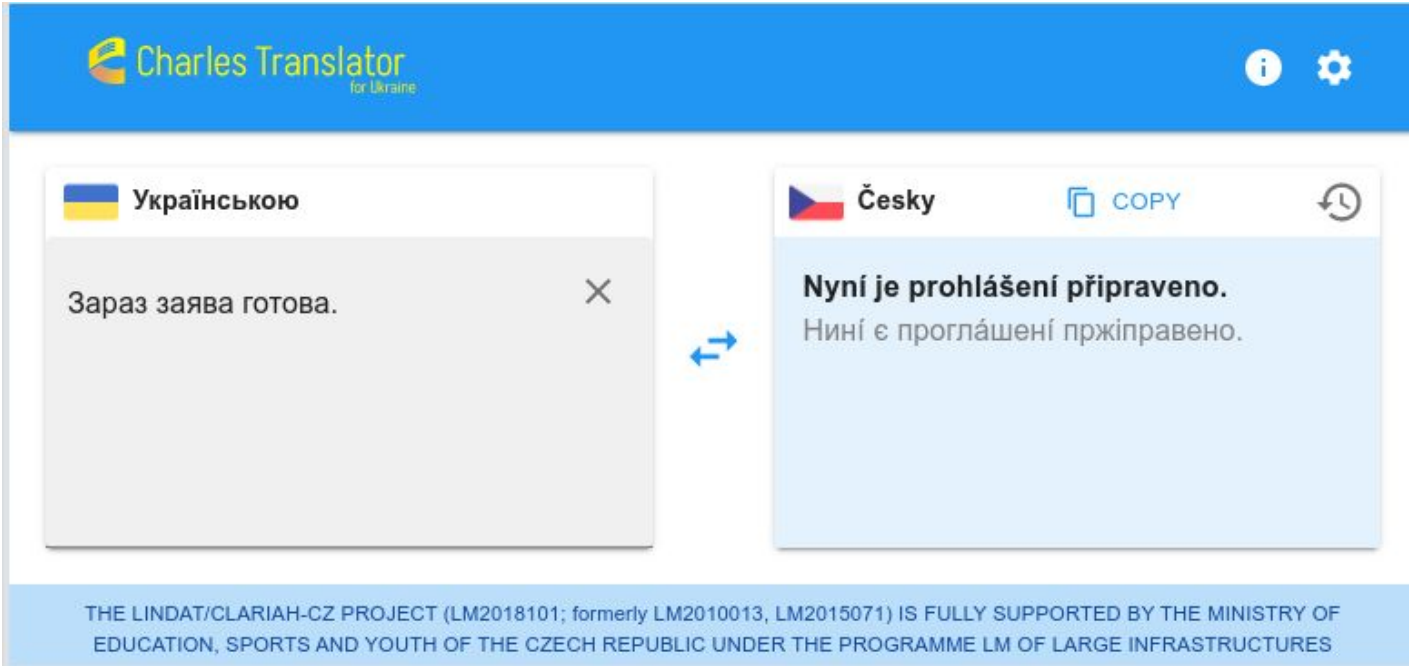
Zkušenosti z Googlu, Microsoftu, LMU München, University of Edinburgh

U4U (UFAL for Ukraine) Timeline

- 1.3. 2022: *Hi folks,*
  Shouldn't we build a Czech-Ukrainian translator and put it on Lindat?
- 12.3. 2022: hackathon na vývoj front-endu

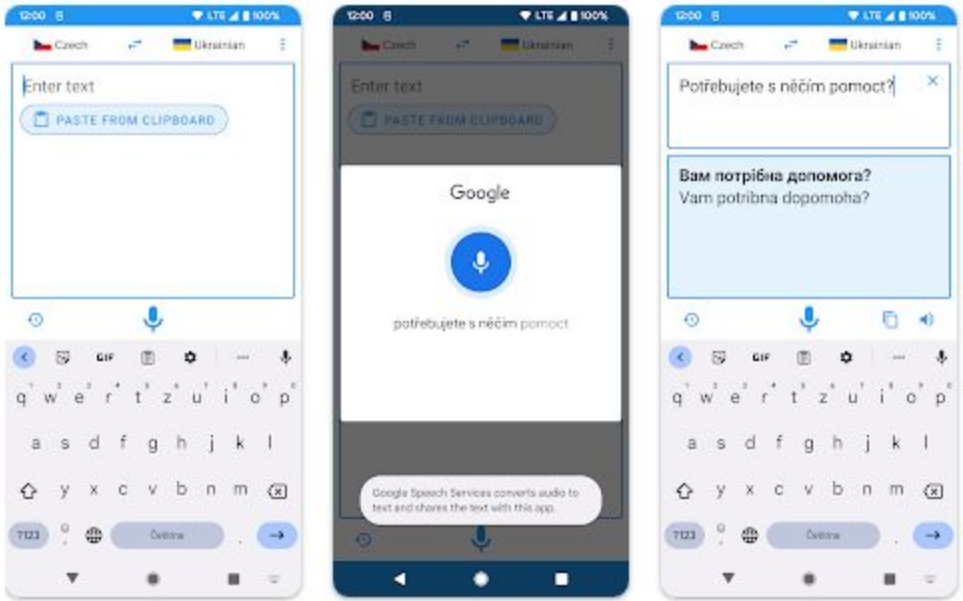
U4U (UFAL for Ukraine) Timeline

- 1.3. 2022: *Hi folks,*
  Shouldn't we build a Czech-Ukrainian translator and put it on Lindat?
- 12.3. 2022: hackathon na vývoj front-endu
- 14.3. 2022: první překladový systém cs ⇆ uk



U4U (UFAL for Ukraine) Timeline

- 1.3. 2022: *Hi folks,*
  Shouldn't we build a Czech-Ukrainian translator and put it on Lindat?
- 12.3. 2022: hackathon na vývoj front-endu
- 14.3. 2022: první překladový systém cs ⇆ uk
- 11.4. 2022: vylepšené modely
- 24.5. 2022: aplikace pro Android

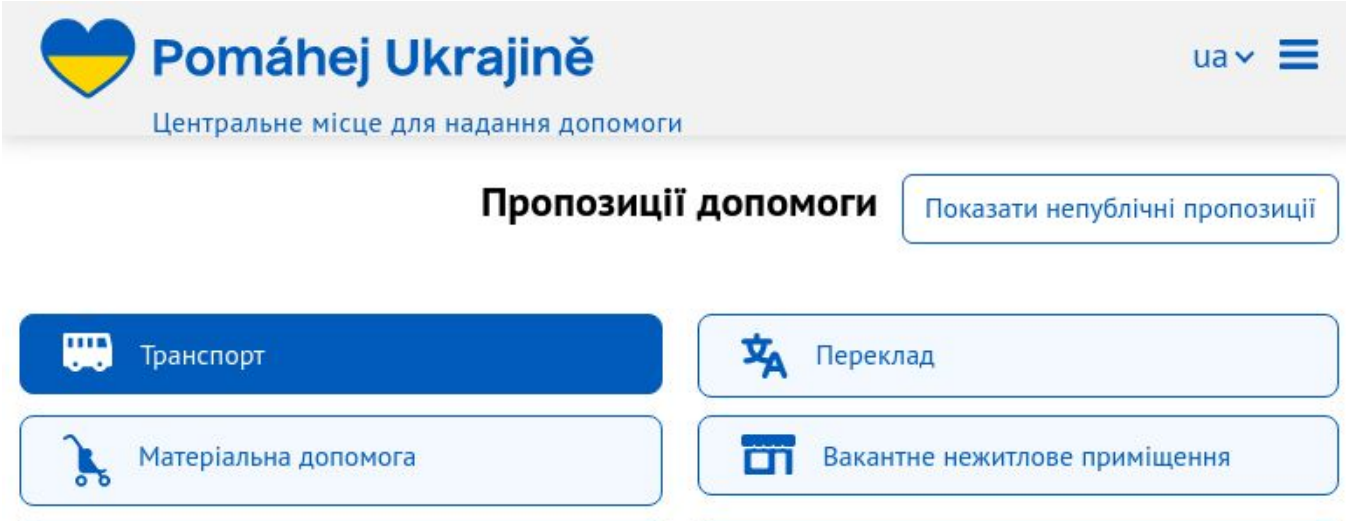


U4U (UFAL for Ukraine) Timeline

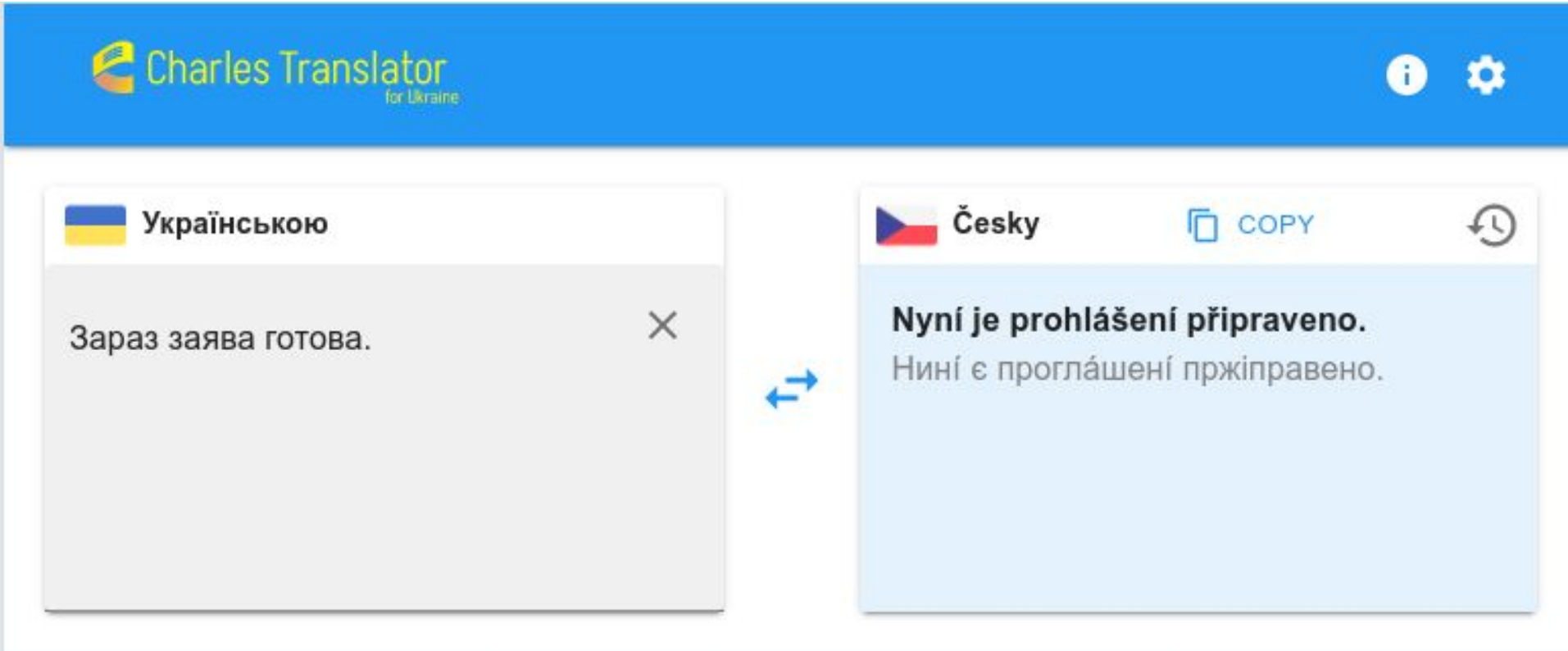
- 1.3. 2022: Hi folks,
- Shouldn't we build a Czech-Ukrainian translator and put it on Lindat?
- 12.3. 2022: hackathon na vývoj front-endu
- 14.3. 2022: první překladový systém cs ⇆ uk
- 11.4. 2022: vylepšené modely
- 24.5. 2022: aplikace pro Android
- Od září: evaluace 11 systémů v mezinárodní soutěži

Why to develop our own MT?

- Quick solution until Big Tech companies improve their Ukrainian models
- Direct translation (not pivoted through English)
- Collect “in-domain” test set
- Focus on the needs of refugees and humanitarian organizations
- Offer free API (and on-premise for sensitive data)



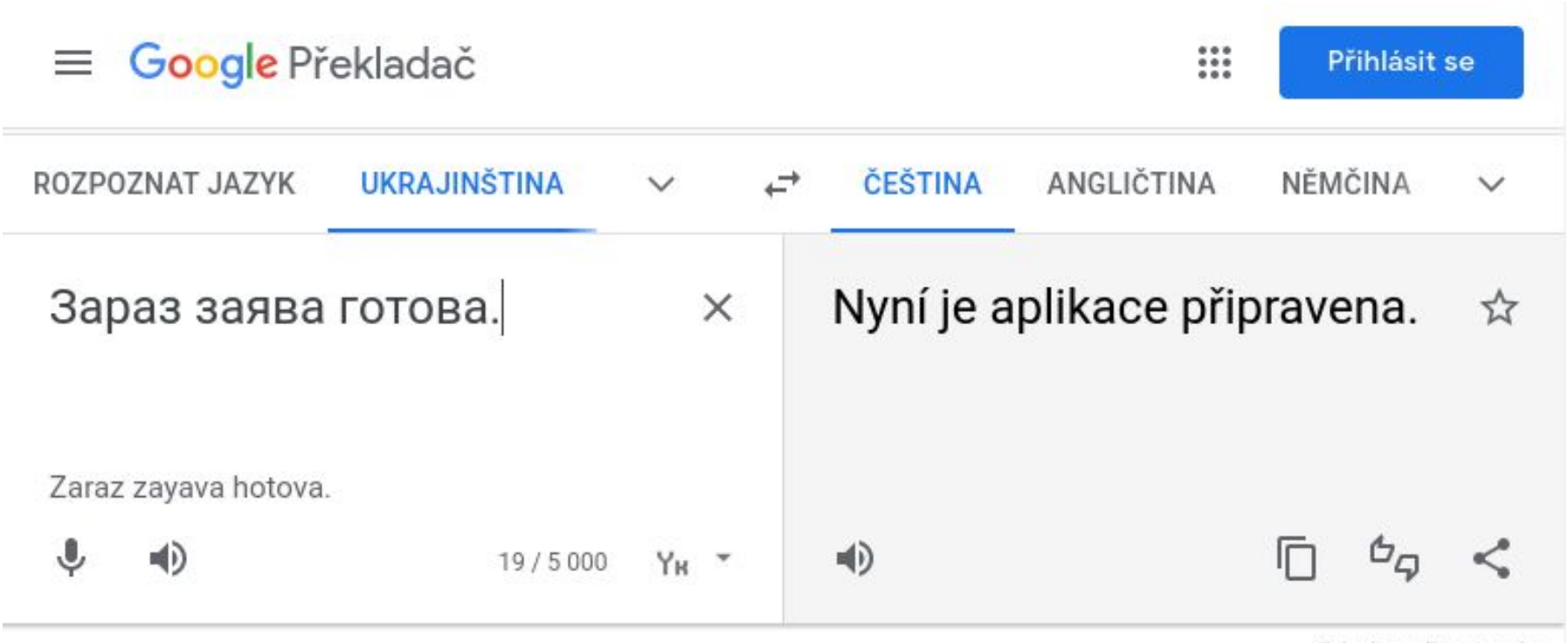
Why direct translation?



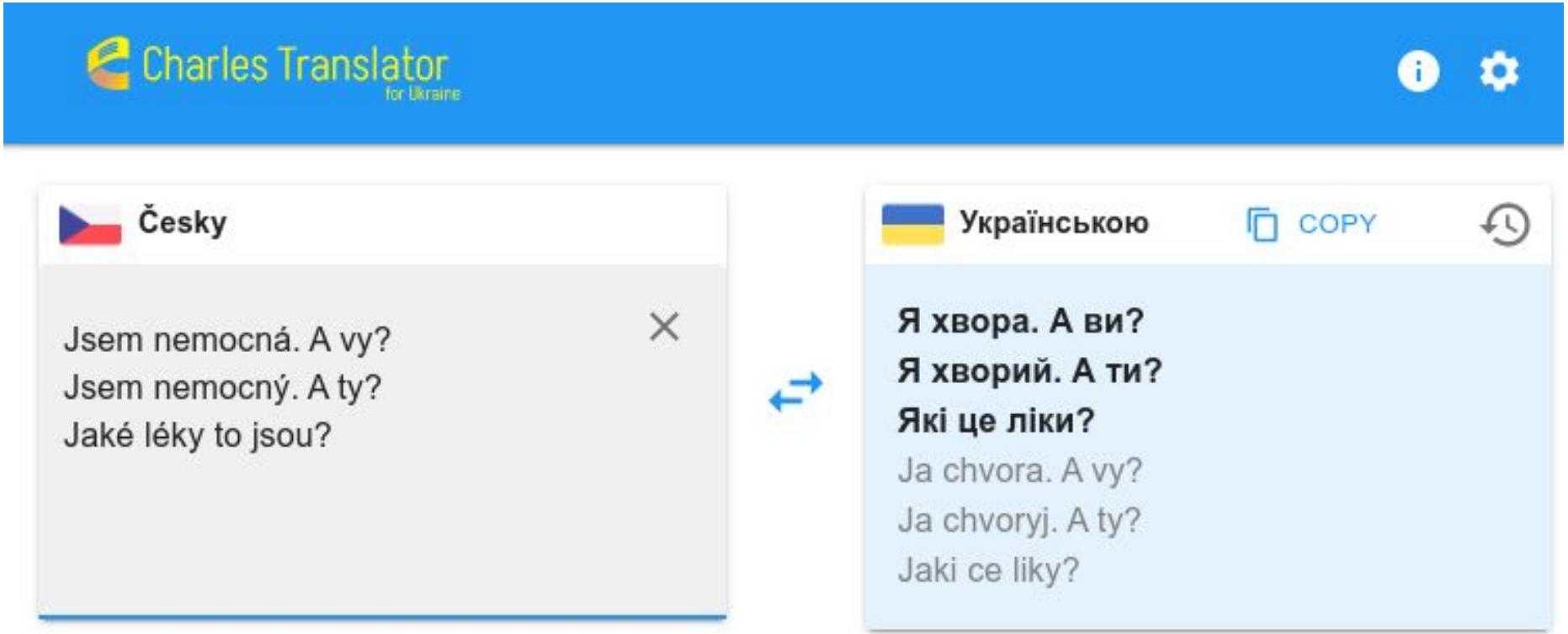
THE LINDAT/CLARIAH-CZ PROJECT (LM2018101; formerly LM2010013, LM2015071) IS FULLY SUPPORTED BY THE MINISTRY OF EDUCATION, SPORTS AND YOUTH OF THE CZECH REPUBLIC UNDER THE PROGRAMME LM OF LARGE INFRASTRUCTURES

Why direct translation?

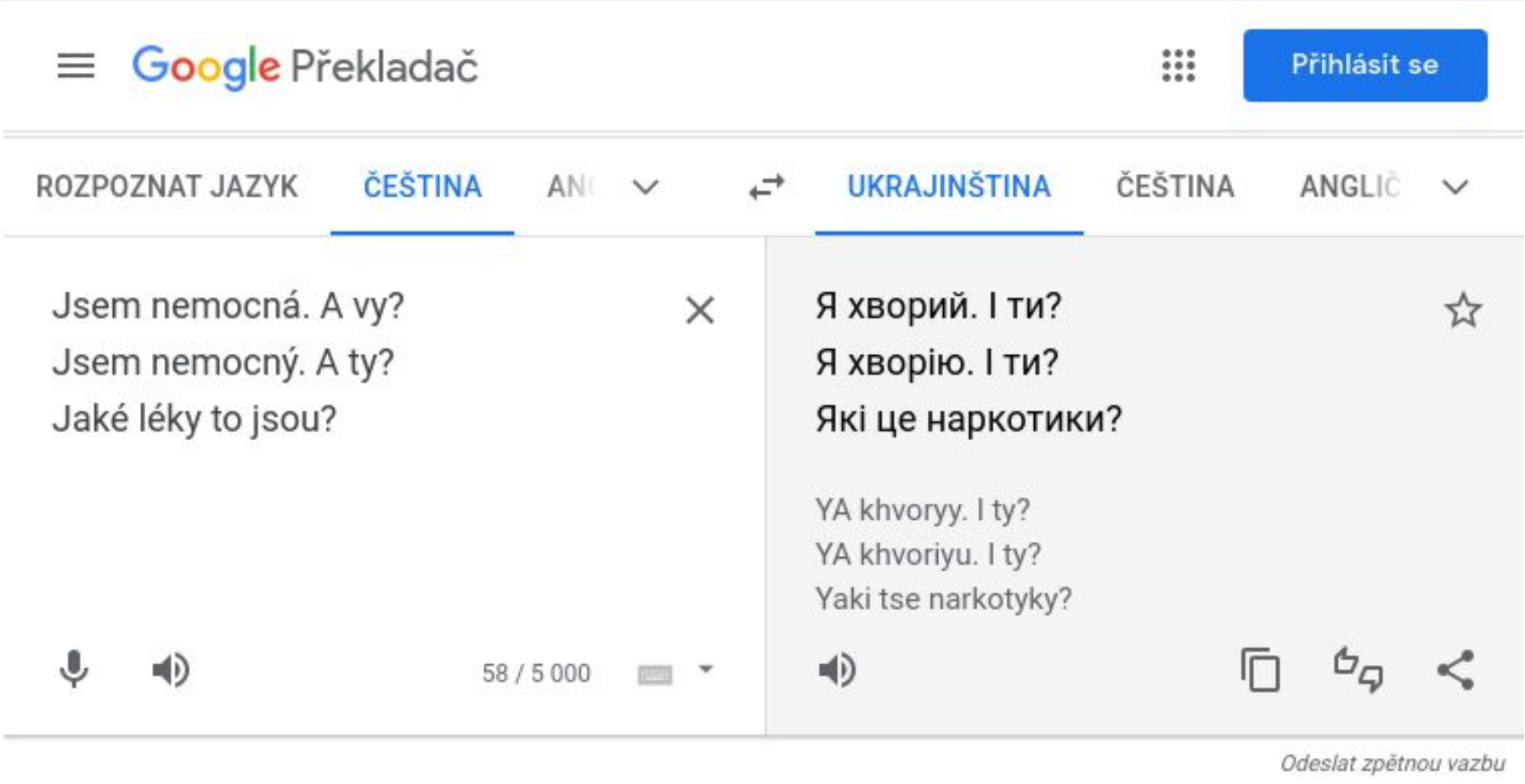
заява ⇒ application ⇒ aplikace



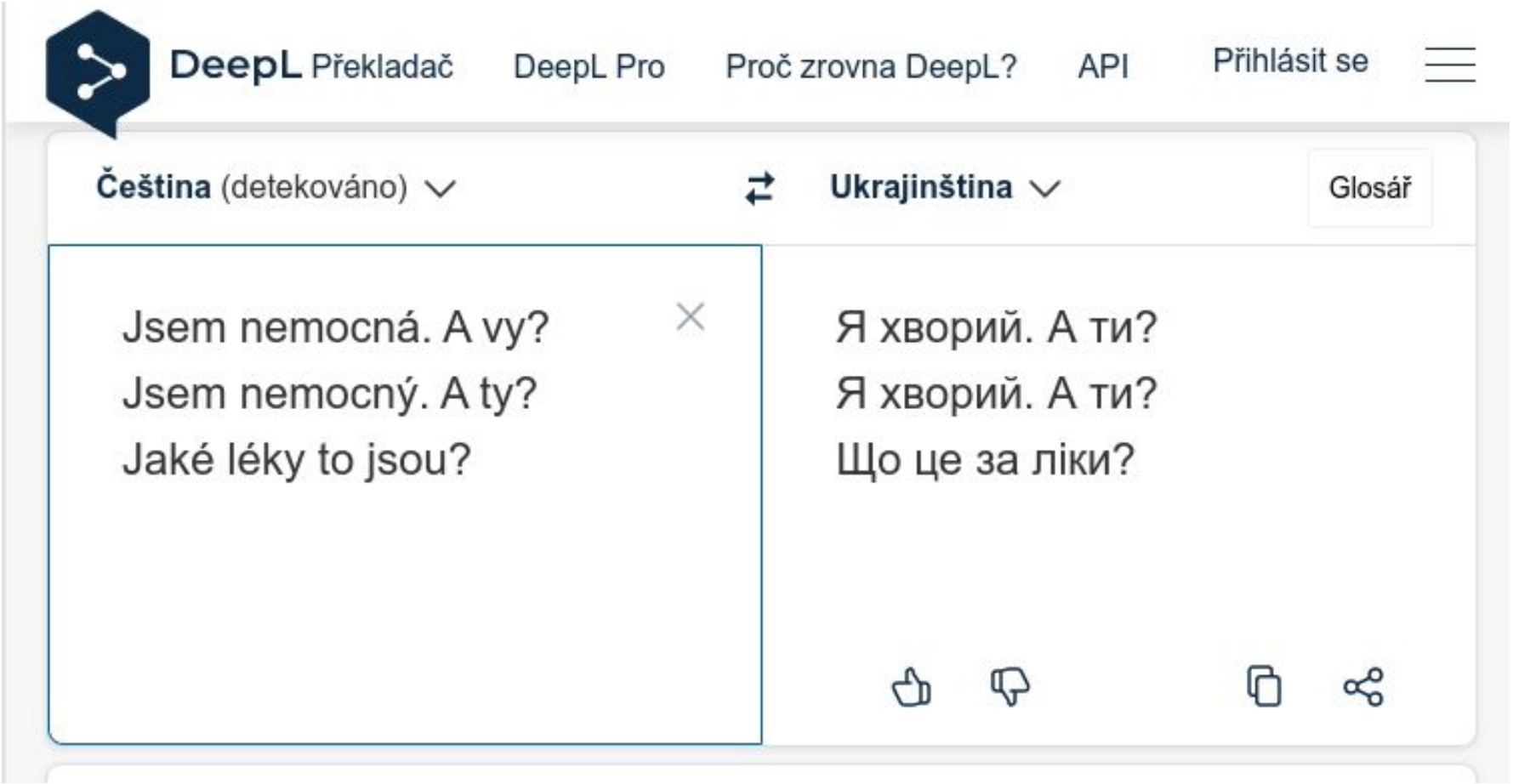
Why direct translation?



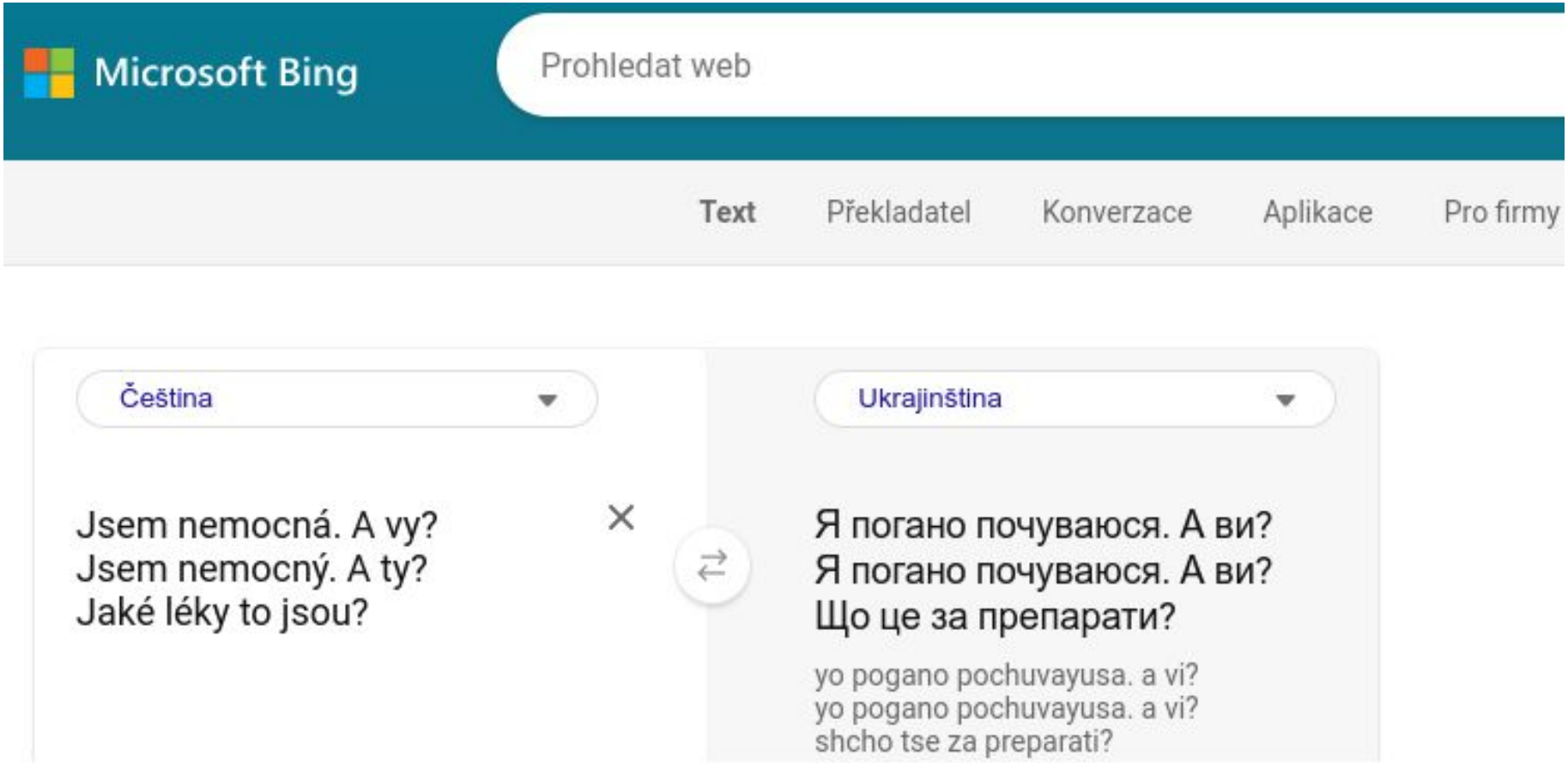
Why direct translation?



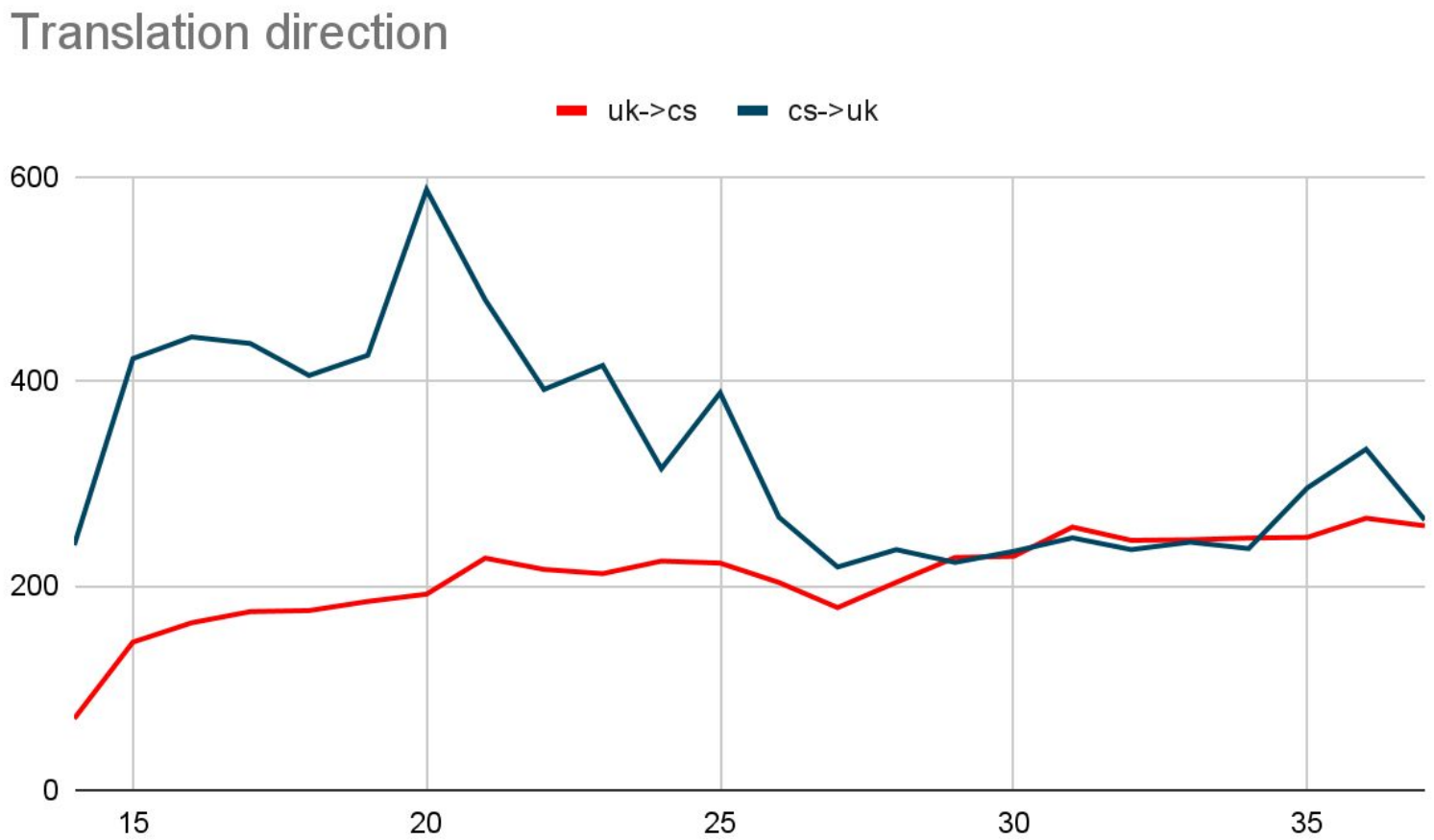
Why direct translation?



Why direct translation?



Thousands of translations per week (April-September)



Dostupné zdroje: paralelní data

Zdroj	Počet vět	Počet vět po filtrování
CCMatrix	3 991 954	3 883 834
MultiCCAligned	1 606 502	1 199 169
fb-WikiMatrix	848 961	809 970
OpenSubtitles	730 804	273 296
QED (educational subtitles)	161 020	138 431
TED talks	115 351	106 003
WikiMatrix	104 983	99 328
intercorp-manualigned	88 533	72 686
KDE4	133 673	63 604
bible-uedin	7 953	7 566
Tatoeba	2 905	2 140
wikimedia	1 959	1 557
EUbookshop	1 506	1 309
Ubuntu	232	178
GNOME	150	81
	7 796 486	**6 659 152**

- Dat je dost -- obejdeme se bez ruštiny = můžeme trénovat přímý model
- Velká část vět ve veřejných datasetech je rusky a ne ukrajinsky
- Paralelní datasety jsou připravované automaticky, párování vět často selže: filtrujeme věty podezřelé délky

Dostupné zdroje: monolinguální data

Využijí se na tzv. back-translation.

Množství není problém, problém je data vybrat podle zaměření překladače!

Čeština:

- zpravodajské texty připravené pro překlad mezi češtinou a angličtinou
- 50M vět, vysoká kvalita

Ukrajinština:

- zpravodajské texty, právnické texty + 6M textů z webu
- nahradíme reprezentativní korpusem 80M vět -- lingvisticky předzpracovaný, chybí mu interpunkce, strojově doplníme

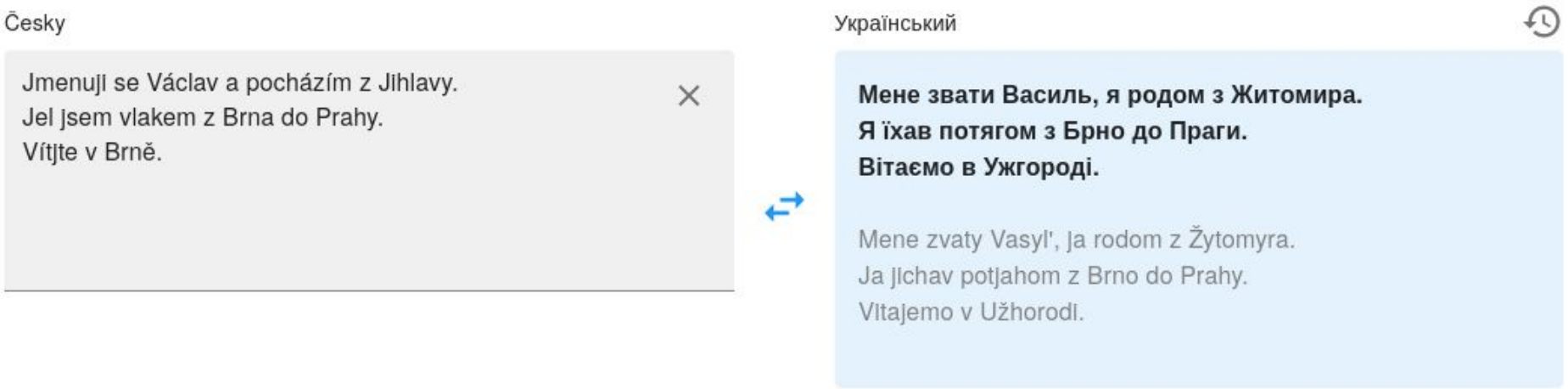
Ukrajinská monolinguální data

	Originally	Filtered
Legal text	7.5 M	7.3 M
News	11.3 M	9.5 M
Web crawl	6.0 M	4.2 M
Total	**24.8 M**	**21.0 M**

Opět byla část ve skutečnosti rusky

Problémy s trénovacími daty

Špatně napárované věty z webu a z Wikipedie…



Problém měření kvality

- Pro výzkumné účely Facebook připravil FLORES 101 test set
  - Věty z Wikipedia pro 101 jazyků
  - Složité věty plné nepravděpodobných termínů
  - Není reprezentativní pro naše použití
- Pracovní testovací data
  - Fráze pro lékaře od MZČR, příručka pro učitele od MŠMT, informace z MVČR, fráze z česko-ukrajinské konverzace
  - Problém: úřady často využívají Google Translate, když to použijeme testování, Google vypadá falešně kvalitní
- Ze skutečného využívání překladače jsme vytvořili realistická testovací data => zapojili jsme do mezinárodní soutěže

Překladač je a bude volně dostpuný

https://translator.cuni.cz

Česko-anglický překlad

https://lindat.mff.cuni.cz/services/translation

Shrnutí

- Umělá inteligence a jazykové technologie jsou jedním z informatických oboru na MFF UK
- Umělá inteligence většinou znamená strojové učení a neuronové sítě
- Strojový překlad se učí z dat pomocí neuronových sítí
- Už nějakou dobu funguje v zásadě dobře, ale stále mnoho problémů
- Ukrajinsko-český překlad dostupný **`translate.cuni.cz`**

`https://ufal.cz`