Strojový překlad pro investigativní účely

Seminář AFCEA, Policejní akademie

Ondřej Bojar, ÚFAL

17. leden 2019

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

Osnova

- Obtížnost strojového překladu (machine translation, MT).
- Strojový překlad před rokem 2016, “frázový”.
- Strojový překlad v současné době
  - = Neuronový ~~neurální~~ strojový překlad (NMT).
    - Jak a co se hluboké neuronové sítě se učí.
- Ukázky výstupů.
- Překlad vietnamských SMS.

Proč je překlad těžký?

- Idiomatická spojení (kick the bucket, beat around the bush).
- Frázová a nepravidelná slovesa (run on, run off, run after).

Proč je strojový překlad těžký?

- Idiomatická spojení (kick the bucket, beat around the bush).
- Frázová a nepravidelná slovesa (run on, run off, run after).

- Víceznačnost a význam slov.
- Cílový slovní tvar.
- Negace.
- Koordinace.
- Zájmena.
- Pořádek slov (tj. i vzdálenost mezi slovy).

Proč je strojový překlad těžký?

- Idiomatická spojení (kick the bucket, beat around the bush).
- Frázová a nepravidelná slovesa (run on, run off, run after).

- **Víceznačnost a význam slov.**
- **Cílový slovní tvar.**
- **Negace.**
- Koordinace.
- Zájmena.
- Pořádek slov (tj. i vzdálenost mezi slovy).

Víceznačnost a význam slov

The plant is next to the bank.

Víceznačnost a význam slov

The plant is next to the bank.
Spal celou Petkevičovu přednášku.
Ženu holí stroj.

Víceznačnost a význam slov

The plant is next to the bank.
Spal celou Petkevičovu přednášku.
Ženu holí stroj.

SRC	One tap and the machine issues a slip with a number.
REF	Jedno ťuknutí a ze stroje vyjede papírek s číslem.
ÚFAL 2011a	Z jednoho kohoutku a stroj vydá složenky s číslem.
ÚFAL 2011b	Jeden úder a stroj vydá složenky s číslem.
Google 2011	Jedním klepnutím a stroj problémy skluzu s číslem.

Víceznačnost a význam slov

The plant is next to the bank.
Spal celou Petkevičovu přednášku.
Ženu holí stroj.

SRC	One tap and the machine issues a slip with a number.
REF	Jedno ťuknutí a ze stroje vyjede papírek s číslem.
ÚFAL 2011a	Z jednoho kohoutku a stroj vydá složenky s číslem.
ÚFAL 2011b	Jeden úder a stroj vydá složenky s číslem.
Google 2011	Jedním klepnutím a stroj problémy skluzu s číslem.
Google 2017–8	Jeden kohoutek a zařízení vydává skluzu s číslem.
ÚFAL 2018	Jedno klepnutí a přístroj vydá lístek s číslem.

Při překladu je nutno pochopit vstup

I	saw	two	green	striped	cats	.

… a možnosti se násobí …

I	saw	two	green	striped	cats	.
já	pila	dva	zelený	pruhovaný	kočky	.
	pily	dvě	zelená	pruhovaná	koček	
	…	dvou	zelené	pruhované	kočkám	
	viděl	dvěma	zelení	pruhovaní	kočkách	
	viděla	dvěmi	zeleného	pruhovaného	kočkami	
	…		zelených	pruhovaných		
	uviděl		zelenému	pruhovanému		
	uviděla		zeleným	pruhovaným		
	…		zelenou	pruhovanou		
	viděl jsem		zelenými	pruhovanými		
	viděla jsem		…	…		

Do češtiny navíc musíme trefit tvar...

I	saw	two	green	striped	cats	.
já	pila	dva	zelený	pruhovaný	**kočky**	.
	pily	**dvě**	zelená	pruhovaná	koček	
	...	dvou	**zelené**	**pruhované**	kočkám	
	viděl	dvěma	zelení	pruhovaní	kočkách	
	viděla	dvěmi	zeleného	pruhovaného	kočkami	
	...		zelených	pruhovaných		
	uviděl		zelenému	pruhovanému		
	uviděla		zeleným	pruhovaným		
	...		zelenou	pruhovanou		
viděl jsem			zelenými	pruhovanými		
viděla jsem			...	...		

… přičemž tvar závisí na kontextu …

I	saw	two	green	striped	cats	.
já	pila	dva	zelený	pruhovaný	**kočky**	.
	pily	**dvě**	zelená	pruhovaná	koček	
	…	**dvou**	**zelené**	**pruhované**	kočkám	
	viděl	dvěma	zelení	pruhovaní	**kočkách**	
	viděla	dvěmi	zeleného	pruhovaného	kočkami	
	…		**zelených**	**pruhovaných**		
			zelenému	pruhovanému		
zrak mi utkvěl na			zeleným	pruhovaným		
	…		zelenou	pruhovanou		
viděl jsem			zelenými	pruhovanými		
viděla jsem			…	…		

Obtížnost vyhodnocování kvality překladu

Kolik je správných překladů následující věty?

And even though he is a political veteran,
the Councilor Karel Brezina responded similarly.

Obtížnost vyhodnocování kvality překladu

Příklady ze 71 tisíc správných překladů anglické věty:

And even though he is a political veteran,
the Councilor Karel Brezina responded similarly.

A ačkoli ho lze považovat za politického veterána, radní Březina reagoval obdobně.
Ač ho můžeme prohlásit za politického veterána, reakce radního Karla Březiny byla velmi obdobná.
A i přestože je politický matador, radní Karel Březina odpověděl podobně.
A přestože je to politický veterán, velmi obdobná byla i reakce radního K. Březiny.
A radní K. Březina odpověděl obdobně, jakkoli je politický veterán.
A třebaže ho můžeme považovat za politického veterána, reakce Karla Březiny byla velmi podobná.
Byť ho lze označit za politického veterána, Karel Březina reagoval podobně.
Byť ho můžeme prohlásit za politického veterána, byla i odpověď K. Březiny velmi podobná.
K. Březina, i když ho lze prohlásit za politického veterána, odpověděl velmi obdobně.
Odpověď Karla Březiny byla podobná, navzdory tomu, že je politickým veteránem.
Radní Březina odpověděl velmi obdobně, navzdory tomu, že ho lze prohlásit za politického veterána.
Reakce K. Březiny, třebaže je politický veterán, byla velmi obdobná.
Velmi obdobná byla i odpověď Karla Březiny, ačkoli ho lze prohlásit za politického veterána.

Invitatio. Einleitung.

M. Veni, Puer! disce Sapere.

P. Quid hoc eſt, *Sapere?*

M. Omnia,

L. Komm her/ Knab! lerne Weißheit.

S. Was iſt das/ Weißheit?

L. Alles/

1: Vezmi dvojice trénovacích vět

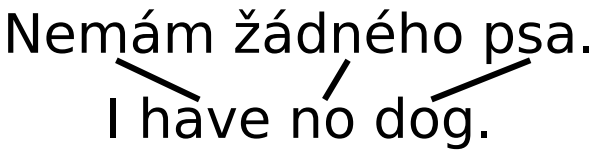
Nemám žádného psa.

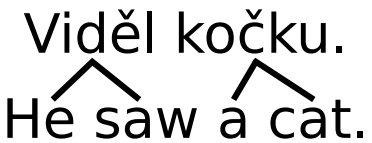
I have no dog.

Viděl kočku.

He saw a cat.

2: Zarovnej k sobě slova





3: Získej překlady “frází”

4: Přijde nový vstup

New input: Nemám kočku.

5: Používej známé fráze...

... I don't have cat.

New input: Nemám kočku.

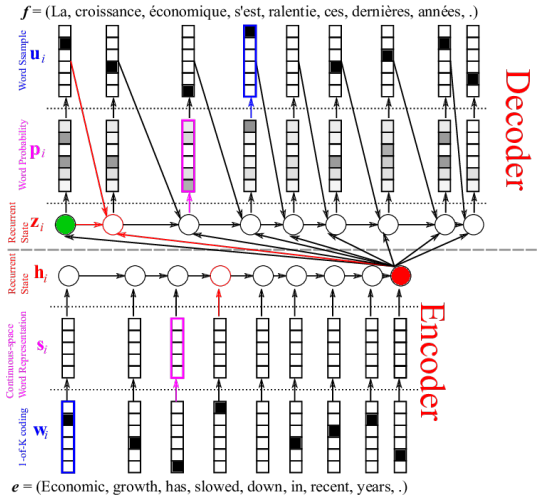
6: ...aby byl výstup plynulý

... I don't have cat.

New input:

7: ...a takhle to dopadlo před rokem 2016.

Neuronový ~~neurální~~ strojový překlad

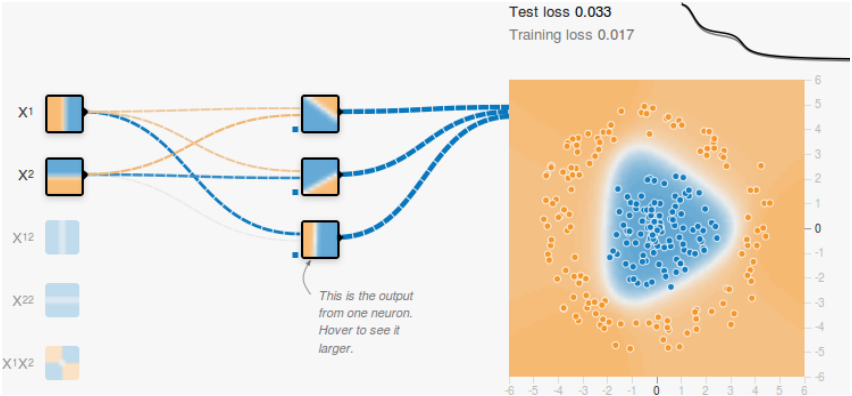


← Model z roku 2014.

- Od té doby tři čtyři generace dalších.

Sutskever et al. (2014);
https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-2/

playground.tensorflow.org

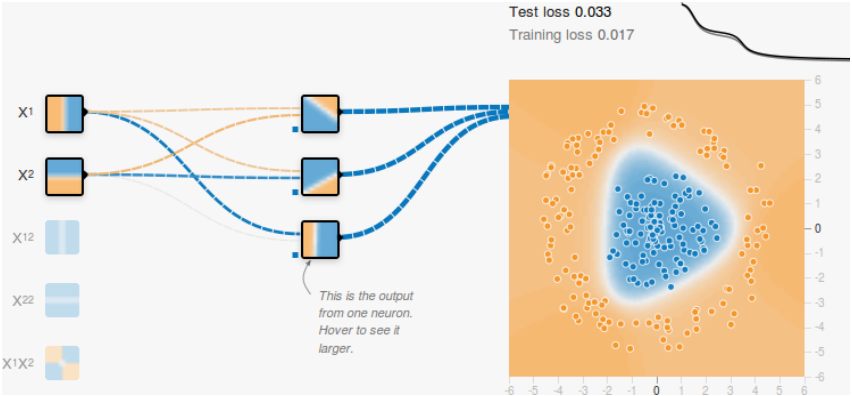


$$-0.43x_1 - 0.89x_2 + 2.0 > 0$$

a $-0.67x_1 + 0.89x_2 + 2.1 > 0$

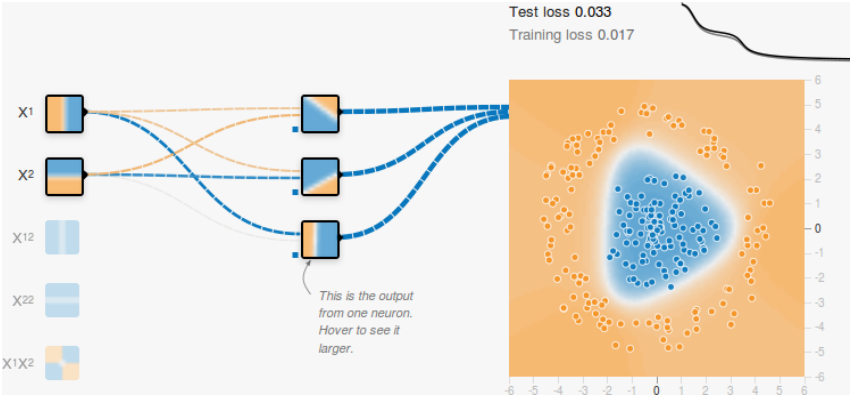
a $1.4x_1 - 0.067x_2 + 2.3 > 0$

“Program” je jen výpočet a test...



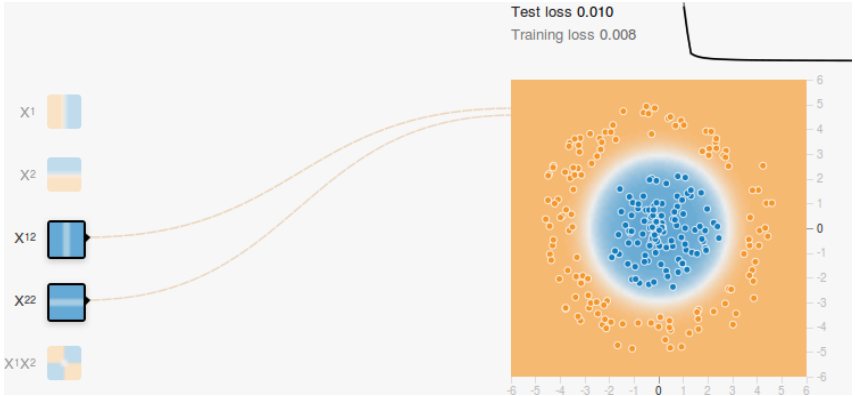
Modrá pokud: $1\tanh(-0.43x_1 - 0.89x_2 + 2.0)$
$+1\tanh(-0.67x_1 + 0.89x_2 + 2.1)$
$+1\tanh(1.4x_1 - 0.067x_2 + 2.3) - \pi/2 > 0$

… s automaticky uhodnutými parametry



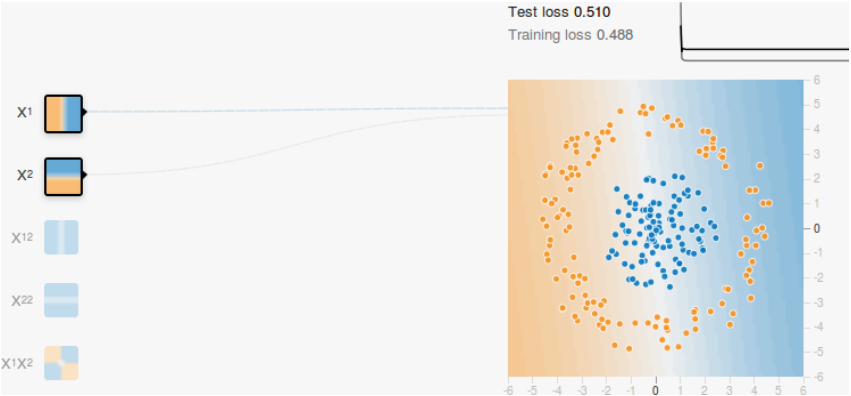
Modrá pokud: $1\tanh(-0.43x_1 - 0.89x_2 + 2.0)$
$+1\tanh(-0.67x_1 + 0.89x_2 + 2.1)$
$+1\tanh(1.4x_1 - 0.067x_2 + 2.3) - \pi/2 > 0$

Ideální vstupy

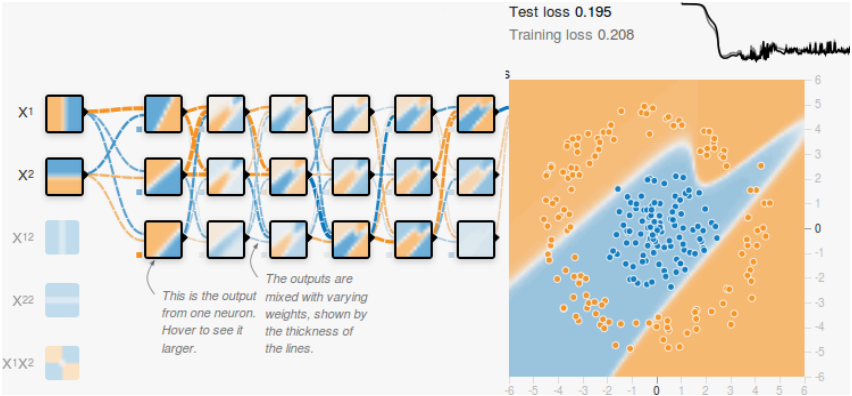


$$1x_1^2 + 1x_2^2 - 1 < 0$$

Nevhodné vstupy a malá hloubka

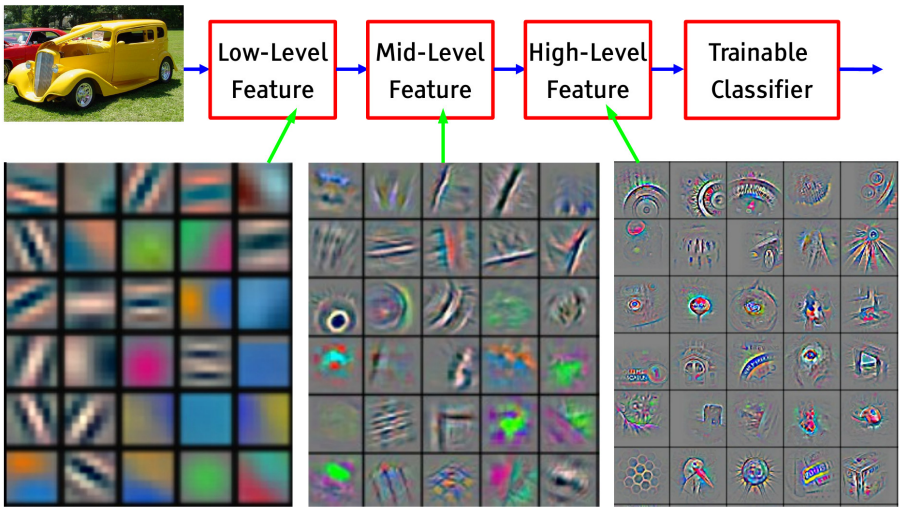


Příliš složitá síť se nenatrénuje



Hluboká síť pro klasifikaci obrázků

- It's deep if it has more than one stage of non-linear feature transformation



Hluboké sítě se učí reprezentaci

Trénování klasifikátoru:

- Na základě trénovacích dat
  (ukázkové vstupy a očekávané výstupy)
- se neuronová síť (neural network, NN) sama naučí
- čeho si ve vstupech všímat.

"**Reprezentace**" je nový souřadný systém.

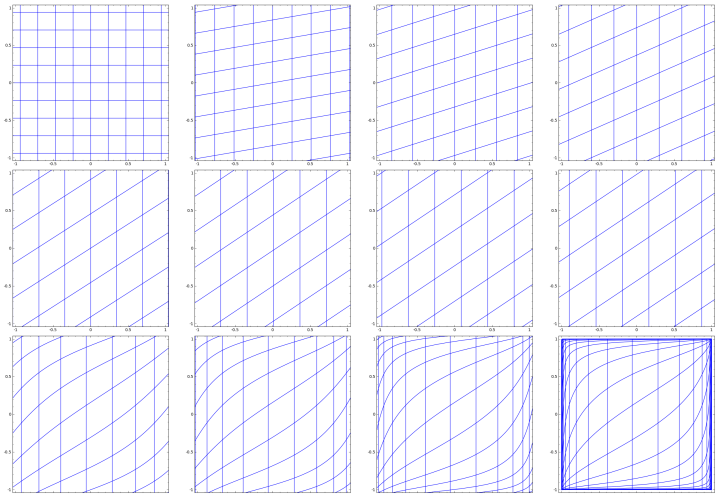
- Místo 3 rozměrů (x, y, barva) dostaneme
- 2000 rozměrů: (slonovitost, počet čápů, modrost, ...)
- nalezených tak, aby nejvíc pomáhaly uhodnout výstup.

Změna souřadnic jednou vrstvou $\tanh(Wx + b)$

Zkosení:
W

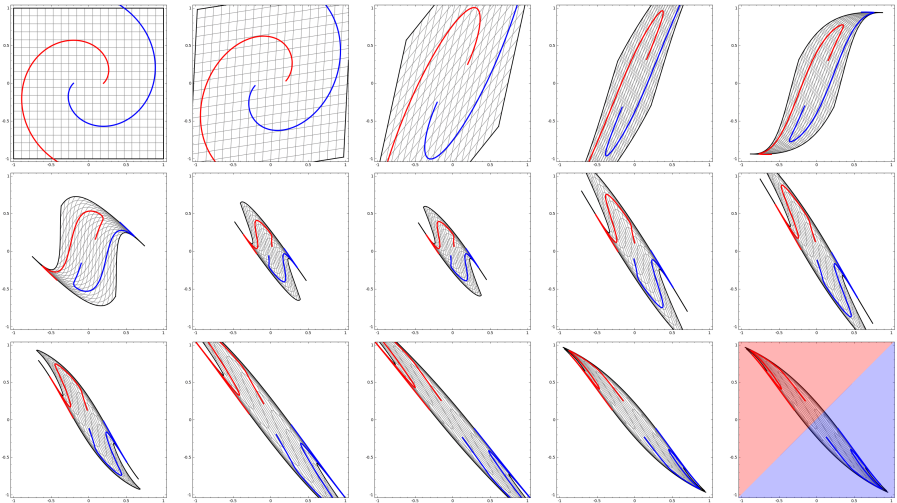
Posun:
b

Nelinearita:
tanh



Animace z `http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/`

Čtyři vrstvy dokáží rozplést spirály



Animace z `http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/`

Jak do NN nalít text?

- Každé slovo zapiš jako vektor nul a jedniček (“1-hot repr.”):

$$\text{cat} \mapsto (0, 0, \ldots, 0, 1, 0, \ldots, 0)$$

- Věta je pak reprezentována úzkou maticí:

		the	cat	is	on	the	mat
↑	a	0	0	0	0	0	0
	about	0	0	0	0	0	0
	...	...	...	...	...	...	...
	cat	0	**1**	0	0	0	0
	...	...	...	...	...	...	...
	is	0	0	**1**	0	0	0
	...	...	...	...	...	...	...
	the	**1**	0	0	0	**1**	0
	...	...	...	...	...	...	...
↓	zebra	0	0	0	0	0	0

Jak do NN nalít text?

- Každé slovo zapiš jako vektor nul a jedniček (“1-hot repr.”):

$$\text{cat} \mapsto (0, 0, \ldots, 0, 1, 0, \ldots, 0)$$

- Věta je pak reprezentována úzkou maticí:

		the	cat	is	on	the	mat
↑	a	0	0	0	0	0	0
	about	0	0	0	0	0	0
	...	...	...	...	...	...	...
Slovník:	cat	0	**1**	0	0	0	0
1.3M angličtina	...	...	...	...	...	...	...
2.2M čeština	is	0	0	**1**	0	0	0
	...	...	...	...	...	...	...
	the	**1**	0	0	0	**1**	0
	...	...	...	...	...	...	...
↓	zebra	0	0	0	0	0	0

Jak do NN nalít text?

- Každé slovo zapiš jako vektor nul a jedniček (“1-hot repr.”):

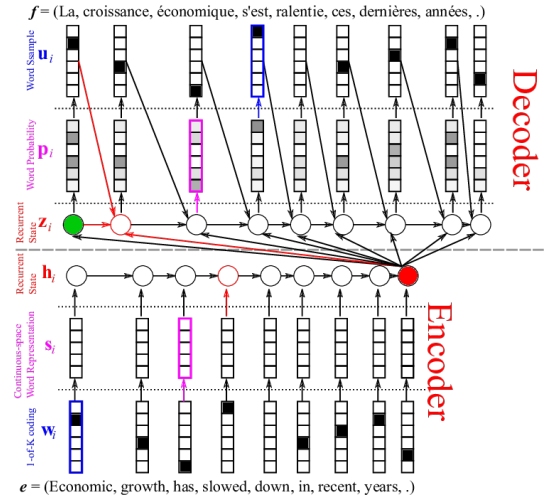
$$\text{cat} \mapsto (0, 0, \ldots, 0, 1, 0, \ldots, 0)$$

- Věta je pak reprezentována úzkou maticí:

		the	cat	is	on	the	mat
↑	a	0	0	0	0	0	0
	about	0	0	0	0	0	0
	...	...	...	...	...	...	...
Slovník:	cat	0	**1**	0	0	0	0
1.3M angličtina	...	...	...	...	...	...	...
2.2M čeština	is	0	0	**1**	0	0	0
	...	...	...	...	...	...	...
	the	**1**	0	0	0	**1**	0
	...	...	...	...	...	...	...
↓	zebra	0	0	0	0	0	0

- Hned v prvním kroku se převede na “embedding”, $\sim$2000 složek.

Neuronový strojový překlad



- One-hot vektor indikuje konkrétní slovo.
- Embedding reprezentuje slova ve spojitém prostoru $\mathbb{R}^n$.
- Enkodér pohlcuje slova a vytváří vektorovou reprezentaci celé věty (●).
- Dekodér z počátečního stavu ● generuje výstupní slova.

Jak byl NMT dobrý v roce 2017?

SRC A 28-year-old chef who had recently moved to San Francisco was found dead in the stairwell of a local mall this week.

Osmadvacetiletý kuchař, který se nedávno přestěhoval do San Francisca, byl tento týden nalezen mrtvý na schodišti místního obchodního centra.

Osmadvacetiletý šéfkuchař, který se nedávno přistěhoval do San Franciska, byl tento týden ∅ schodech místního obchodu.

Výstupy nejlepšího systému v soutěži: `http://www.statmt.org/wmt17`

Jak byl NMT dobrý v roce 2017?

SRC	A 28-year-old chef who had recently moved to San Francisco was found dead in the stairwell of a local mall this week.
MT	Osmadvacetiletý kuchař, který se nedávno přestěhoval do San Francisca, byl tento týden nalezen mrtvý na schodišti místního obchodního centra.
REF	Osmadvacetiletý šéfkuchař, který se nedávno přistěhoval do San Franciska, byl tento týden ∅ schodech místního obchodu.

Výstupy nejlepšího systému v soutěži: `http://www.statmt.org/wmt17`

Jak byl NMT dobrý v roce 2017?

SRC	A 28-year-old chef who had recently moved to San Francisco was found dead in the stairwell of a local mall this week.
MT	Osmadvacetiletý kuchař, který se nedávno přestěhoval do San Francisca, byl tento týden nalezen mrtvý na schodišti místního obchodního centra.
REF	Osmadvacetiletý šéfkuchař, který se nedávno přistěhoval do San Franciska, byl tento týden ∅ schodech místního obchodu.
SRC	There were creative differences on the set and a disagreement.
	Došlo ke vzniku kreativních rozdílů na scéně a k neshodám. Na place byly tvůrčí rozdíly a neshody.

Výstupy nejlepšího systému v soutěži: `http://www.statmt.org/wmt17`

Jak byl NMT dobrý v roce 2017?

SRC	A 28-year-old chef who had recently moved to San Francisco was found dead in the stairwell of a local mall this week.
MT	Osmadvacetiletý kuchař, který se nedávno přestěhoval do San Francisca, byl tento týden nalezen mrtvý na schodišti místního obchodního centra.
REF	Osmadvacetiletý šéfkuchař, který se nedávno přistěhoval do San Franciska, byl tento týden ∅ schodech místního obchodu.

SRC	There were creative differences on the set and a disagreement.
REF	Došlo ke vzniku kreativních rozdílů na scéně a k neshodám.
MT	Na place byly tvůrčí rozdíly a neshody.

Výstupy nejlepšího systému v soutěži: `http://www.statmt.org/wmt17`

Jak dobrý je NMT dnes?

	Ave. %	Ave. z	System
1	84.4	0.667	CUNI-TRANSFORMER
2	79.8	0.521	UEDIN
	78.6	0.483	**Referenční překlad**
4	68.1	0.128	ONLINE-B
5	59.4	−0.178	ONLINE-A
6	54.1	−0.354	ONLINE-G

Varování:

- Lidé překládali celé dokumenty, MT jednotlivé věty.
- Vyhodnocování probíhalo po jednotlivých větách.

MT stále dělá vážné chyby (vítěz 2017)

SRC ... Frank initially stayed in hostels...
MT ... Frank původně zůstal v **Budějovicích**...

SRC Most of the Clintons' income...
MT Většinu příjmů **Kliniky**...

SRC The 63-year-old has now been made a special representative...
MT 63letý **mladík** se nyní stal zvláštním zástupcem...

SRC He listened to the moving stories of the women.
MT Naslouchal **pohyblivým** příběhům žen.

...a někdy katastrofické chyby

SRC Criminal Minds star Thomas Gibson sacked after hitting producer
REF Thomas Gibson, hvězda seriálu Myšlenky zločince, byl propuštěn po té, co uhodil režiséra
MT **Kriminalisté Minsku** hvězdu Thomase Gibsona **vyhostili** po **zásahu** producenta

SRC ...add to that its long-standing grudge...
REF ...přidejte k tomu svou dlouholetou nenávist...
MT ...přidejte k tomu svou dlouholetou **záštitu**...
(grudge → zášť → záštita)

Tragicky selhávají ale i lidé

Tragicky selhávají ale i lidé

Google Translate

English | Spanish | French | Welsh - detected

Nid wyf yn y swyddfa ar hyn o bryd.
Anfonfwch unrhyw waith i'w gyfieithu.

73/5000

English | Hebrew | Czech

Translate

V současné době nejsem v kanceláři. Zašlete prosím jakoukoli práci, kterou chcete přeložit.

Překlad vietnamských SMS

- Díky podpoře z OP Praha pól růstu adaptujeme MT pro NPC PČR.
- Cíl našeho podprojektu (2017–2018):

  In-house překlad vietnamských SMS.

Specifika:

- Relativně málo studovaná dvojice jazyků.
- 1.1M paralelních vět mimo doménu (filmové titulky, TED talks).
- Pouze 55 tisíc vietnamských SMS.
- Nutnost pracovat výhradně na pracovišti NPC PČR.
- ... a navíc neumíme vietnamsky.

Vyhodnocení nejlepších konfigurací

- Automaticky proti referenčnímu překladu (BLEU skóre):

	Veřejný test set	Test set PČR
Frázový překlad	**5.50**	2.73
Neuronový překlad	3.74	**9.94**

- Ruční vyhodnocení na vzorku 50 vět test setu PČR:

	Jeden výstup lepší			Oba stejně		Reference
	*	*?	*!	dobré	špatné	nepoužitelná
Frázový	0	1	0	1	8	10
Neuronový	20	7	3			

Počet vět: * správně, *? s nepřesností, *! s hrubou chybou významu.

Inherentní neurčitost ve vstupu

- Podobně jako krátké věty, SMS nenesou kontext pro lepší zjednoznačnění slov.

Ve vietnamštině komplikováno:

- Chybějící diakritikou: "bo" může být:
  - bộ (jít) • bò (hovězí) • bơ (máslo).
- Chudou morfologií vietnamštiny:
  "Cho anh ta 500 đô la." může znamenat:
  - Dal jsem mu 500 dolarů. • Předej mu 500 dolarů.
  - Předal jsi mu 500 dolarů?

  Vhodný překlad je "předání 500 dolarů jemu" nebo "dárek $500".
  Příklad z dat: "À dang o nha a." je přeloženo jako:
  - Jsem u sebe doma. • Jsi doma?

Minulé a běžící spolupráce

- Série EU projektů o překladu (2006–2018):

  ↳ Vlastní překladač Evropské komise MT@EC, eTranslation.

- Rutinní překlad pro IBM Česká republika:
  - angličtina v kombinaci s češtinou, maďarštinou, japonštinou, arabštinou.
- Další drobné spolupráce, “strojový překlad na zakázku”.

Startující projekty

ELITR: European Live Translator (2019–2021; http://elitr.eu/)

- Strojový překlad a tlumočení mezi všemi jazyky EU a EUROSAI.

Technologie	**Hlavní cíl**	**Pokryto**	**Experimentálně**
Transkripce řeči	En, De	Fr, Sp, It, Ru	Cs
Překlad/tlumočení	{ En, De } → { En, De, Cs }	všechny jazyky EU → všechny jazyky EU	všechny jazyky EUROSAI → všechny jazyky EUROSAI
Sumarizace	angličtina, čeština	–	–

Bergamot: Browser translation (2019–2021)

- Offline překlad (ve webovém prohlížeči).

GAČR CZECH SCIENCE FOUNDATION **NEUREM**^{3}: Neural representations (2019–2023)

- Základní výzkum: trénování na neanotovaných datech; studium, co se sítě učí.

Shrnutí

- Strojový překlad je těžký.
  - Pro různé účely stačí ale různě kvalitní výstupy.
  - Naše hlavní zaměření bylo na produkci dobrých vět.
  - Překlad pro účely vyhledávání by si vystačil se starou technologií.
- Hluboké neuronové sítě často zvládají překlad skvěle.
  - … I překladatelé ostatně dělají chyby.
  - … Navíc hrozí, že vynesou důvěrné informace.
- Při dostatku trénovacích dat na jazycích nezáleží.

References

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. pages 3104–3112.