Promises and Pitfalls of Neural MT

SCIC Lunchtime Session

Ondřej Bojar

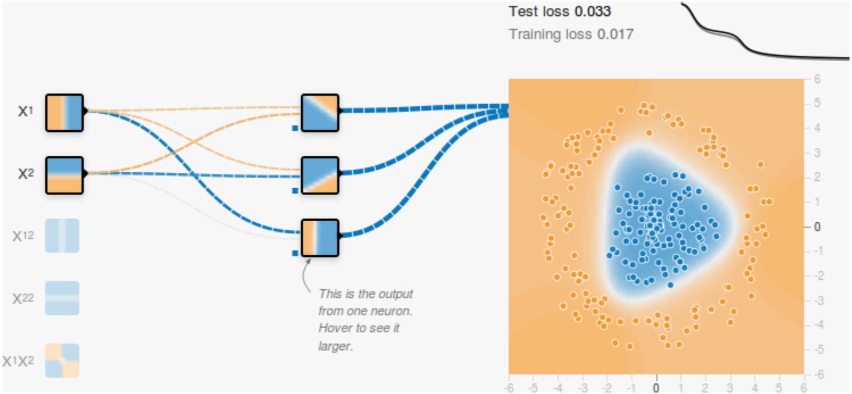
November 11, 2018

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

Outline

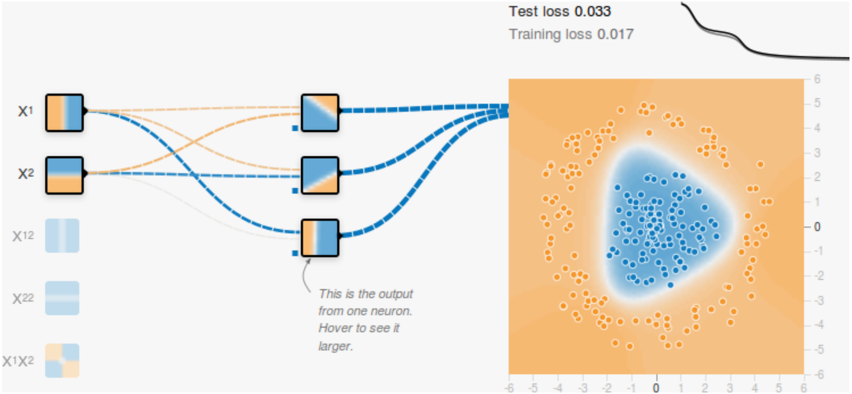
- Neural Networks using High-School Maths Illustrations.
    - The Hype/Hope of Representation Learning.
- Neural Machine Translation (NMT).
    - Overview.
    - Good and Bad Outputs.
    - Is NMT Understanding?
    - Reaching Human Translator Quality.
- ELITR Goals.

playground.tensorflow.org



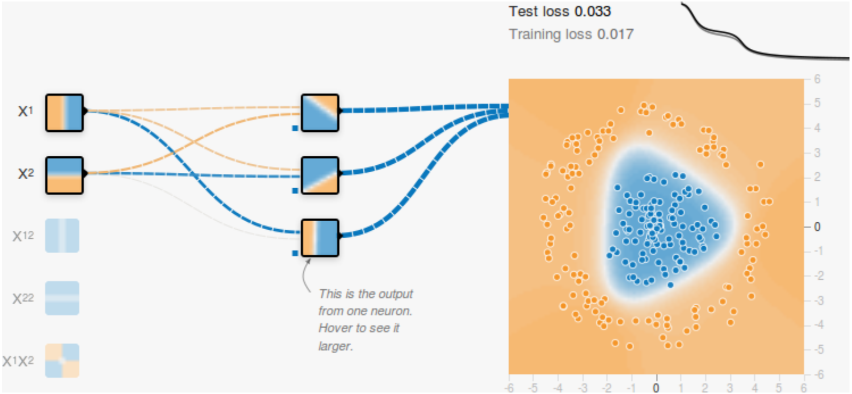
$$-0.43x_1 - 0.89x_2 + 2.0 > 0$$
$$\text{and } -0.67x_1 + 0.89x_2 + 2.1 > 0$$
$$\text{and } 1.4x_1 - 0.067x_2 + 2.3 > 0$$

The “Program” Is Just a Computation...



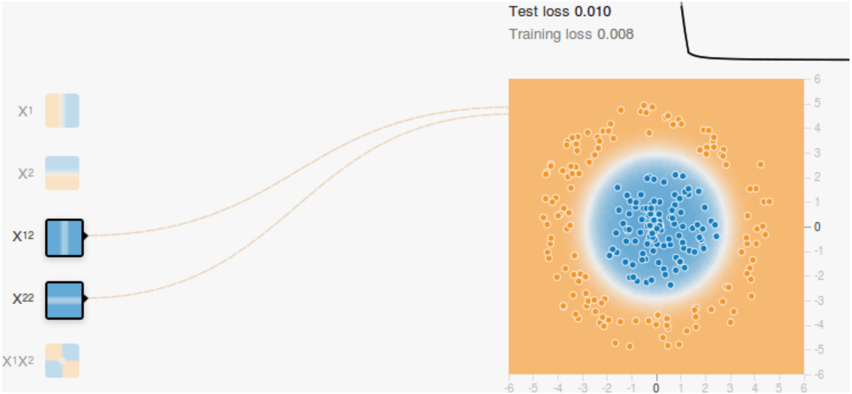
$$\text{In fact: } 1\tanh(-0.43x_1 - 0.89x_2 + 2.0) + 1\tanh(-0.67x_1 + 0.89x_2 + 2.1) + 1\tanh(1.4x_1 - 0.067x_2 + 2.3) - \pi/2 > 0$$

... with Parameters Guessed Automatically



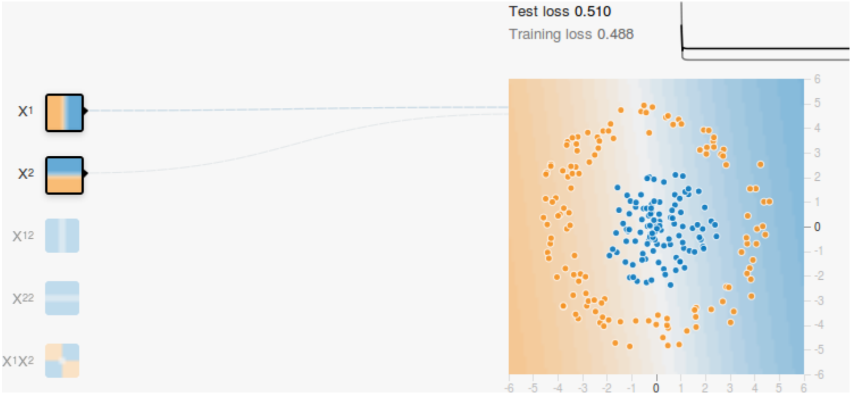
In fact: $\color{red}{1}\tanh(\color{red}{-0.43}x_1 \color{red}{-0.89}x_2 + \color{red}{2.0})$
$+\color{red}{1}\tanh(\color{red}{-0.67}x_1 \color{red}{+0.89}x_2 + \color{red}{2.1})$
$+\color{red}{1}\tanh(\color{red}{1.4}x_1 \color{red}{-0.067}x_2 + \color{red}{2.3}) \color{red}{-\pi/2} > 0$

Perfect Features

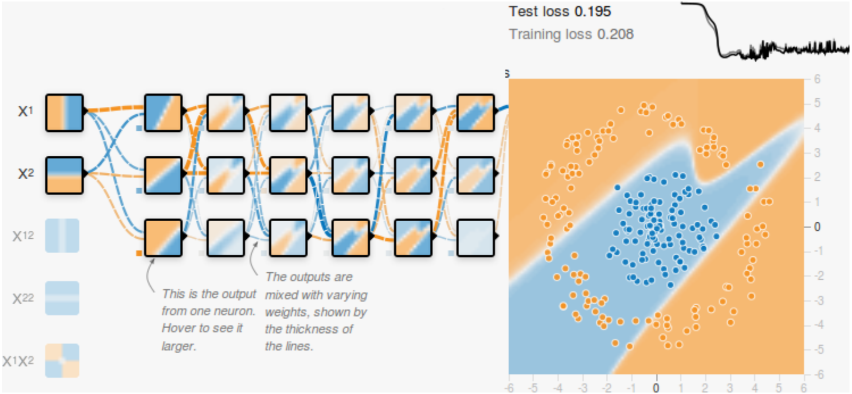


$$1x_1^2 + 1x_2^2 - 1 < 0$$

Bad Features & Low Depth

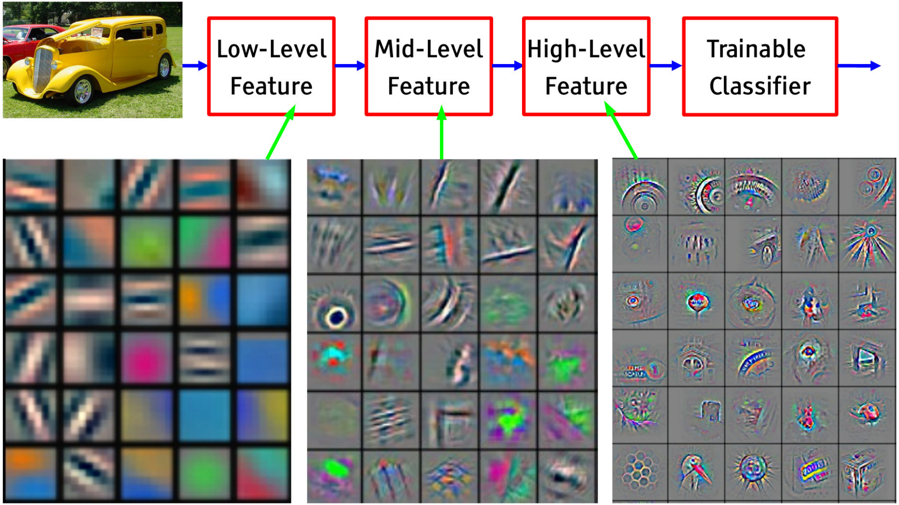


Too Complex NN Fails to Learn



Deep NNs for Image Classification

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

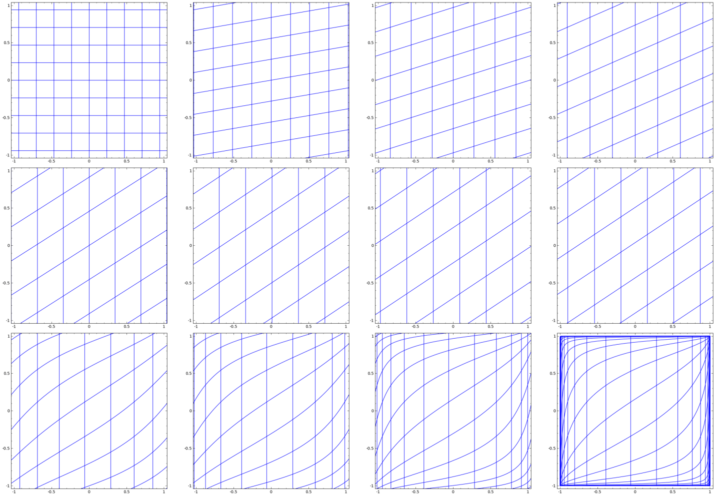


One Layer $\tanh(Wx + b)$, 2D→2D

Skew:
W

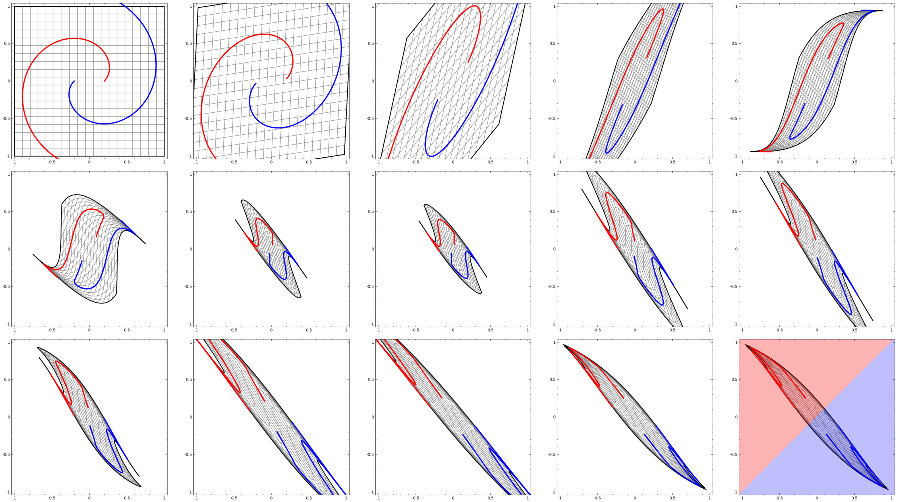
Transpose:
b

Non-lin.:
tanh



Animation by `http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/`

Four Layers, Disentagling Spirals



Animation by `http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/`

Representation Learning and Classification

Representation Learning

- Deep learning can find useful features.
- We can think of these features as new coordinates.

$\Rightarrow$ NNs are learning how to represent the input to make it linearly separable.

Supervised Classification

- A large number of labelled training examples needed.
  - ASR: Speech and transcript.
  - MT: Translated texts.
  - Summarization: Long and abridged text.
- Network trained preferably end-to-end.
  - Errors cummulate if trained piecewise.

Machine Translation as Classification

- In MT, the input and output are **sequences**.
- Sequence-to-sequence as classification:
    - Predict the next word given input and output so far:

$$p(e_1^I|f_1^J) = p(e_1, e_2, \dots e_I|f_1^J) = \prod_{i=1}^{I} p(e_i|e_1, \dots e_{i-1}, f_1^J)$$

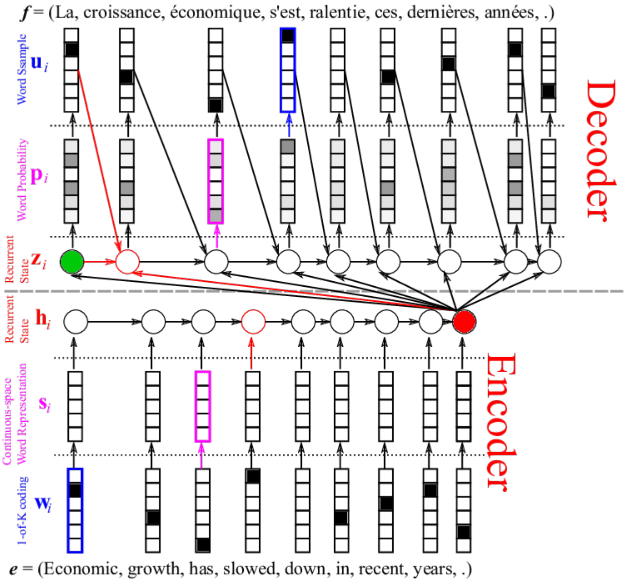
Problem of target vocabulary size:

- 1-2 million **word types** seen in large parallel corpora.

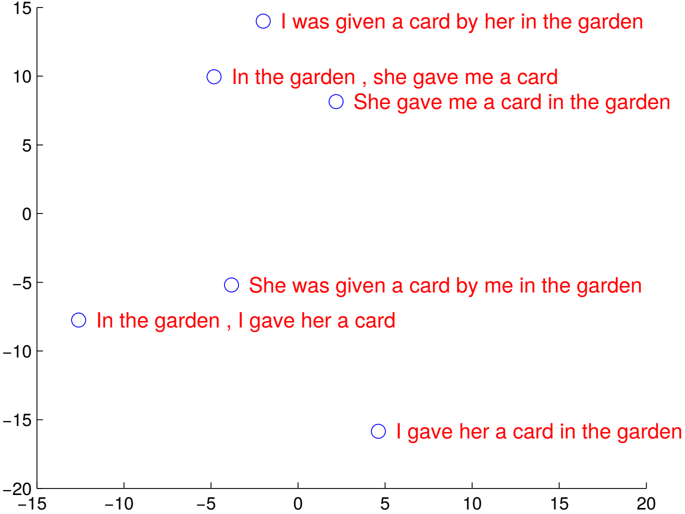
  nejneobhodpodařovávatelnějšími, Donaudampfschifffahrtsgesellschaftskapitän
- NMT can handle only 30–80k dictionaries.
- Resort to subword units (Sennrich et al., 2016):

Orig	český politik svezl migranty
BPE 30k	český politik s@@ vez@@ l mi@@ granty

Encoder-Decoder Architecture

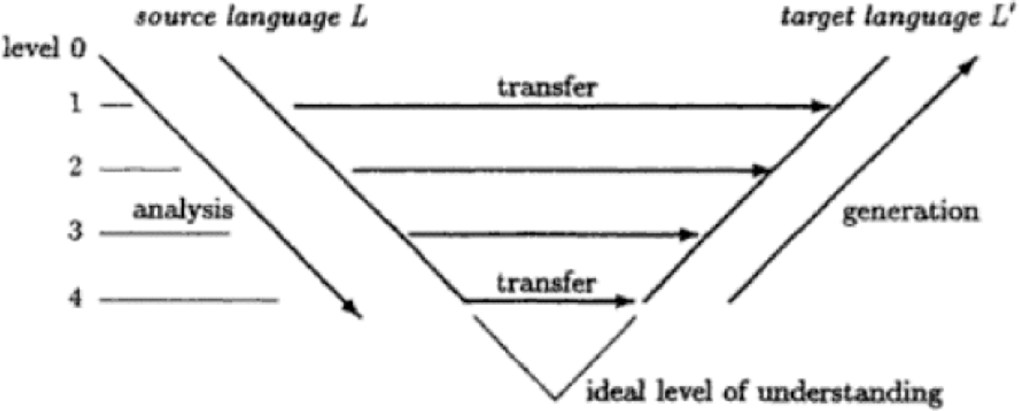


Continuous Space of Sentences



2-D PCA projection of 8000-D space representing sentences (Sutskever et al., 2014).

Interlingua?

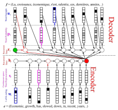


From Vauquois (1968), reproduced by Adam Lopez.

Interlingua?

- Theoretically, a very inspiring concept.
- Need for $2N$ instead of n^2 systems.
- Sceptical view:
  - Need to capture all distinctions in word meanings: `https://en.wikipedia.org/wiki/Eskimo_words_for_snow`
  - Text form underspecifies the meaning, formally captured content underspecifies the form (Lampert, 2001).
  - Interannotator agreement decreases as we proceed along layers of linguistic analysis (Dorr et al., 2010).

Promising Results: Google Interlingua



... simply feed seq2seq architecture with various language pairs.

Source Sent 1 (De)	**2en** versetzen Sie sich mal in meine Lage !
Target Sent 1 (En)	put yourselves in my position .
Source Sent 2 (En)	**2nl** I flew on Air Force Two for eight years .
Target Sent 2 (Nl)	ik heb acht jaar lang met de Air Force Two gevlogen .

- The model of the same size will learn both pairs.
- Hopefully benefiting from various similarities.

(Johnson et al., 2016)

Promising Results: Google Interlingua

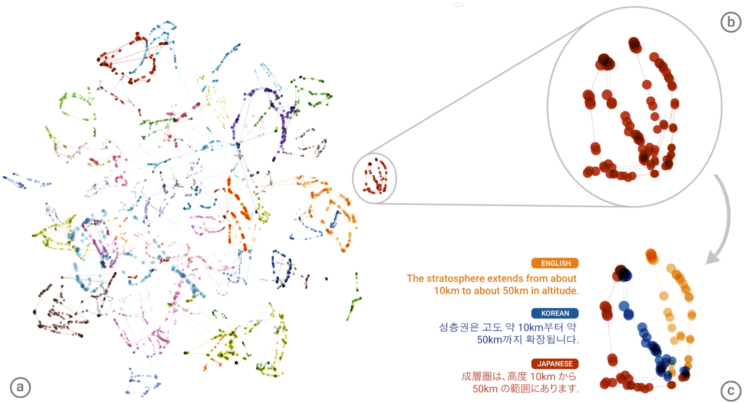
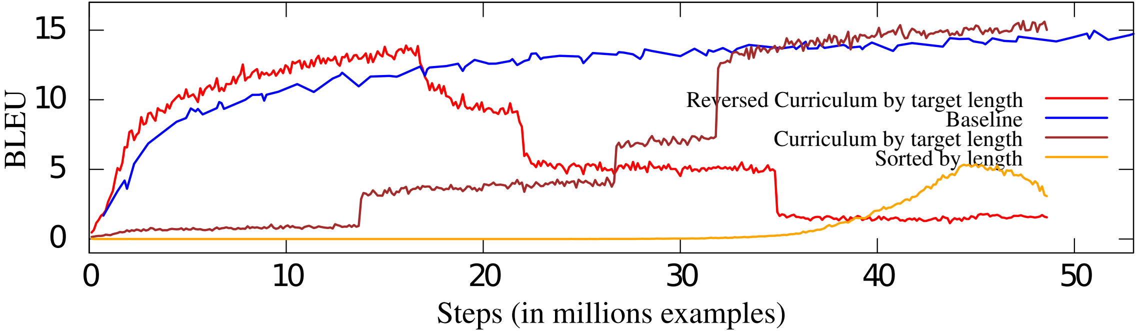


Figure 2: A t-SNE projection of the embedding of 74 semantically identical sentences translated across all 6 possible directions, yielding a total of 9,978 steps (dots in the image), from the model trained on English↔Japanese and English↔Korean examples. (a) A bird's-eye view of the embedding, coloring by the index of the semantic sentence. Well-def ned clusters each having a single color are apparent. (b) A zoomed in view of one of the clusters with the same coloring. All of the sentences within this cluster are translations of "The stratosphere extends from about 10km to about 50km in altitude." (c) The same cluster colored by source language. All three source languages can be seen within this cluster.

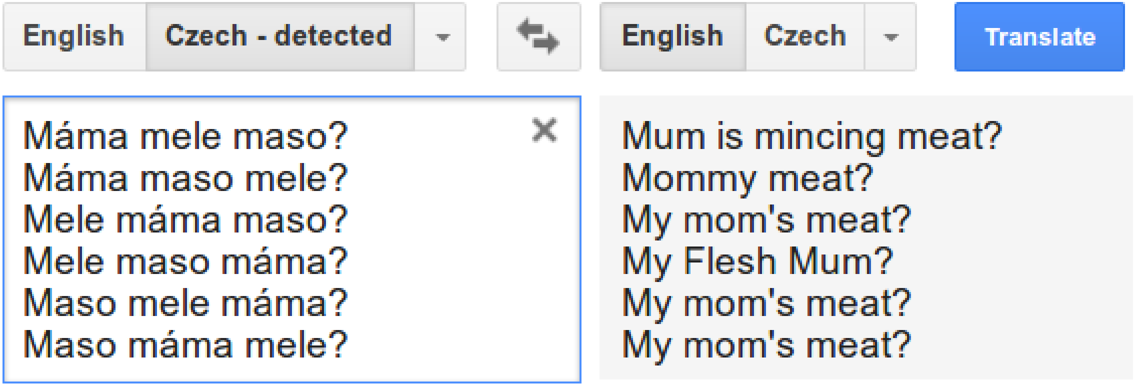
Counterexamples: Catastrophic Forgetting

- Kocmi and Bojar (2017) explore curriculum learning:
  - Start with simpler sentences first, add complex ones later.
- When "simpler" mean "shorter":
  - Clear jumps in score as bins of longer sentences are allowed.
  - Reversed curriculum unlearns to produce long sentences.



Counterexamples: Meaning Not Captured

- Deep learning researchers easily claim that NNs learn the meaning of the sentences.
- This is possible, but not achieved in practice, yet:



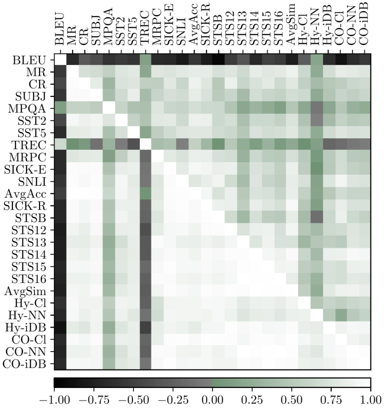
Is NMT Understanding Better?

In Cífka and Bojar (2018), we tested several NMT setups:

- in terms of translation quality,
- in terms of "meaning representation":

Name	Cl.	Task	Example Input and Label
MR	2	sentiment (movies)	*an idealistic love story that brings out the latent 15-year-old romantic in everyone.* (+)
CR	2	product review polarity	*no way to contact their customer service.* (−)
SUBJ	2	subjectivity	*a little weak – and it isn't that funny.* (subjective)
MPQA	2	opinion polarity	*was hoping* (+), *breach of the very constitution* (−)
SST2	2	sentiment (movies)	*contains very few laughs and even less surprises* (−)
SST5	5	sentiment (movies)	*it's worth taking the kids to.* (4)
TREC	6	question type	*What was Einstein s IQ?* (NUM)
MRPC	2	semantic equivalence	*Lawtey is not the first faith-based program in Florida's prison system. / But Lawtey is the first entire prison to take that path.* (−)
SNLI	3	natural language inference	*Two doctors perform surgery on patient. / Two surgeons are having lunch.* (contradiction)
SICK-E	3	natural language inference	*A group of people is near the ocean / A crowd of people is near the water* (entailment)

The Better the MT, the Worse Meaning Repr



Average correlations:

- Semantic tasks against each other:
  - 0.78 ± 0.32 (en→cs)
  - 0.62 ± 0.23 (en→de)
- Translation quality (BLEU) against semantic tasks:
  - −0.57 ± 0.31 (en→cs)
  - −0.54 ± 0.27 (en→de)

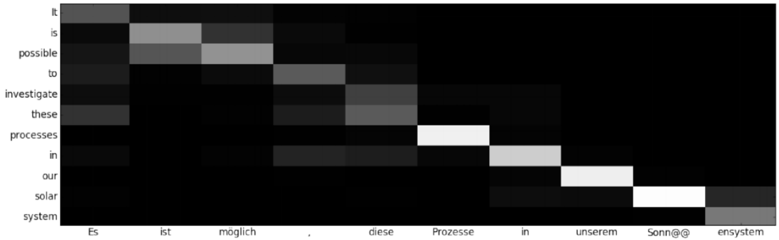
This can be a sign of:

a) Very immature evaluation of meaning.

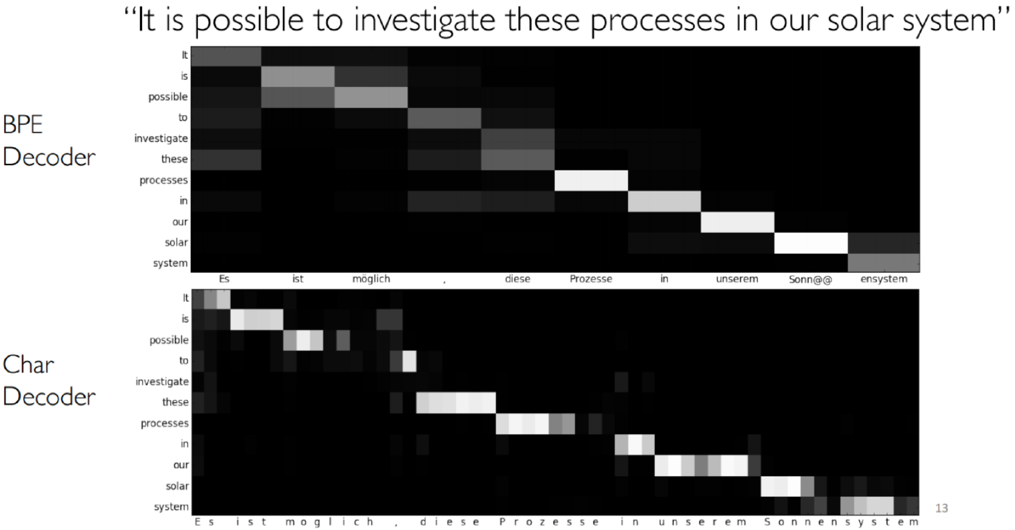
b) NMT deliberately avoiding the meaning.

NMT with Attention

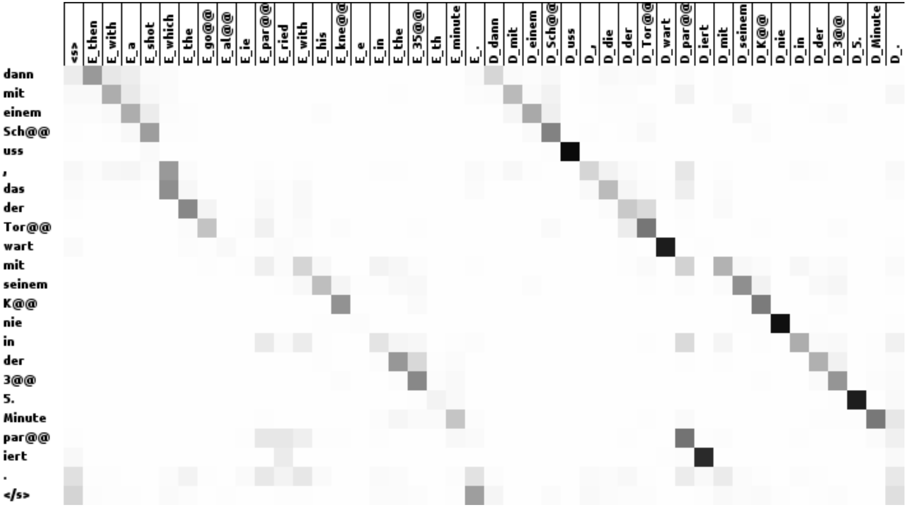
- One fixed-size vector does not capture long sentences well.
- Current NMT systems rely on attention:
  - Store some state after each input token.
  - "Attend" to a mix of all these states at every output step.
- Attention ≈ Alignment, if we collect attention across time.
  - Each column corresponds to one decoder time step.
  - Source tokens correspond to rows.



Attending to Characters

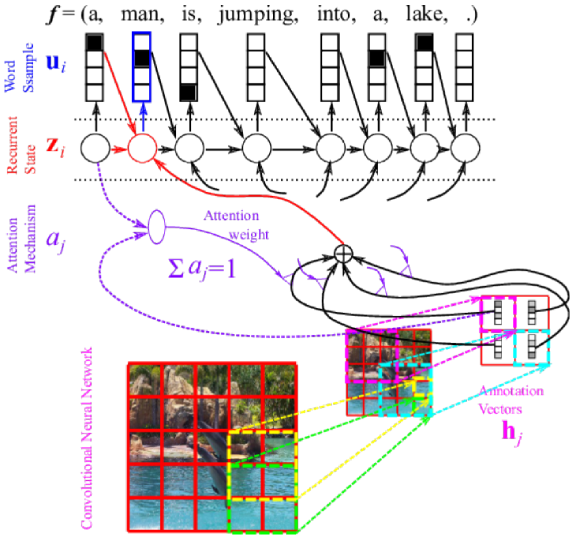


Attending to Two Inputs at Once



Niehues et al. (2016) pre-translates with PBMT and then attends both.

Attending to Image Parts



How Good NMT Was in 2017? (1/2)

SRC	28-Year-Old Chef Found Dead at San Francisco Mall
	28letý šéfkuchař Found Dead v San Francisco Mall
	Osmadvacetiletý šéfkuchař nalezen mrtev v obchodě v San Francisku

How Good NMT Was in 2017? (1/2)

SRC	28-Year-Old Chef Found Dead at San Francisco Mall
MT	28letý šéfkuchař Found Dead v San Francisco Mall
REF	Osmadvacetiletý šéfkuchař nalezen mrtev v obchodě v San Francisku

How Good NMT Was in 2017? (1/2)

SRC	28-Year-Old Chef Found Dead at San Francisco Mall
MT	28letý šéfkuchař Found Dead v San Francisco Mall
REF	Osmadvacetiletý šéfkuchař nalezen mrtev v obchodě v San Francisku

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 $\emptyset$ schodech místního obchodu.

How Good NMT Was in 2017? (1/2)

SRC	28-Year-Old Chef Found Dead at San Francisco Mall
MT	28letý šéfkuchař Found Dead v San Francisco Mall
REF	Osmadvacetiletý šéfkuchař nalezen mrtev v obchodě v San Francisku

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.

How Good NMT Was in 2017? (2/2)

SRC	A spokesperson for Sons & Daughters **said** they were ”shocked and devastated” by his death.
	Mluvčí společnosti Sons & Daughters **uvedla**, že jsou jeho smrtí ”šokováni a zdrceni”.
	Mluvčí restaurace Sons & Daughters **řekl**, že jsou jeho smrtí „šokováni a zničeni“.

How Good NMT Was in 2017? (2/2)

SRC	A spokesperson for Sons & Daughters **said** they were "shocked and devastated" by his death.
MT	Mluvčí společnosti Sons & Daughters **uvedla**, že jsou jeho smrtí "šokováni a zdrceni".
REF	Mluvčí restaurace Sons & Daughters **řekl**, že jsou jeho smrtí „šokováni a zničeni“.

How Good NMT Was in 2017? (2/2)

SRC	A spokesperson for Sons & Daughters **said** they were ”shocked and devastated” by his death.
MT	Mluvčí společnosti Sons & Daughters **uvedla**, že jsou jeho smrtí ”šokováni a zdrceni”.
REF	Mluvčí restaurace Sons & Daughters **řekl**, že jsou jeho smrtí „šokováni a zničeni“.

SRC	”He found an apartment, he was dating a girl,” Louis Galici**a** told KGO.
	„Našel si byt, chodil s dívkou,“ řekl Louis Galici**a** pro KGO.
	”Našel si byt, chodil s holkou,” řekl Louis Galici**e** KGO.

How Good NMT Was in 2017? (2/2)

SRC	A spokesperson for Sons & Daughters **said** they were "shocked and devastated" by his death.
MT	Mluvčí společnosti Sons & Daughters **uvedla**, že jsou jeho smrtí "šokováni a zdrceni".
REF	Mluvčí restaurace Sons & Daughters **řekl**, že jsou jeho smrtí „šokováni a zničeni“.

SRC	"He found an apartment, he was dating a girl," Louis Galicia told KGO.
REF	„Našel si byt, chodil s dívkou,“ řekl Louis Galicia pro KGO.
MT	"Našel si byt, chodil s holkou," řekl Louis Galicie KGO.

Luckily ;-), Catastropic Errors Happen

The exact same WMT17 UEDIN MT system produced also:

SRC	... said Frank initially stayed in **hostels**...
MT	... řekl, že Frank původně zůstal v **Budějovicích**...
↳*Gloss*	... said that Frank initically stayed in **Budweis**...
SRC	Most of the **Clintons'** income...
MT	Většinu příjmů **Kliniky**...
↳*Gloss*	Most of the income of the **Clinic**...
SRC	The 63-year-old has now been made a special representative...
MT	63letý **mladík** se nyní stal zvláštním zástupcem...
↳*Gloss*	The 63-year-old **youngster** has now become a special representat

Catastropic Errors Happen (2/2)

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
↳*Gloss*	**Minsk criminal investigators** have **expelled** the star Thomas Gibson after **striking** the producer

SRC	...add to that its long-standing grudge...
REF	...přidejte k tomu svou dlouholetou nenávist...
MT	...přidejte k tomu svou dlouholetou **záštitu**...
↳*Gloss*	...add to that its long-standing **auspices**... (grudge = zášť $\rightarrow$ záštita = auspices)

UEDIN at WMT17

- Our small annotation of up to 185 sentences.
- Blind mix: reference or MT.

Real MT was assumed to be:

	OB	DM	DV
MT	142 (76.8 %)	86 (77.5 %)	72 (87.8 %)
didn't know	34 (18.4 %)	9 (8.1 %)	6 (7.3 %)
human	9 (4.9 %)	16 (14.4 %)	4 (4.9 %)
Total	185 (100.0 %)	111 (100.0 %)	82 (100.0 %)

$\Rightarrow$ 10–20% of outputs indistinguishable from humans.

UEDIN at WMT17

- Our small annotation of up to 185 sentences.
- Blind mix: reference or MT.

Real MT was assumed to be:

	OB	DM	DV
MT	142 (76.8 %)	86 (77.5 %)	72 (87.8 %)
didn't know	34 (18.4 %)	9 (8.1 %)	6 (7.3 %)
human	9 (4.9 %)	16 (14.4 %)	4 (4.9 %)
Total	185 (100.0 %)	111 (100.0 %)	82 (100.0 %)

$\Rightarrow$ 10–20% of outputs indistinguishable from humans.

	OB	DM	DV
almost flawless	17 (9.19 %)	2 (1.80 %)	0 (0 %)
flawless	82 (44.32 %)	37 (33.33 %)	27 (32.93 %)

$\Rightarrow$ 30–50% of outputs flawless or almost flawless.

How Good NMT is Today? (1/2)

Bilingual manual assessment of translation quality in WMT18:

1/10 blocks, 10 items left in block | WMT18SrcDA #1505:Segment #1487 | English → Czech (čeština)

Costs are mounting in the case, with hundreds of pages of affidavits, emails and reports by companies including Deloitte, Pitcher Partners and Charter Keck Cramer filed and top barristers including Allan Myers, QC, and senior solicitors retained by both sides.

— Reference text

V tomto případě rostou Chile má deset členů a koncem srpna by měl společností, včetně společností Deloitte, Pitcher Partners a Charty Keck Cramer, a špičkových obhájců včetně Allana Myerse, QC a vyšších právních zástupců, které si ponechaly obě strany.

— Candidate translation

— How accurately does the above candidate text convey the original semantics of the source text? Slider ranges from Not at all (left) to Perfectly (right).

Reset

Submit

How Good NMT is Today? (2/2)

	Ave. %	Ave. z	System
1	84.4	0.667	CUNI-Transformer
2	79.8	0.521	uedin
	78.6	0.483	**newstest2018-ref**
4	68.1	0.128	online-B
5	59.4	−0.178	online-A
6	54.1	−0.354	online-G

Caveats:

- Humans translated whole documents, MT individual segments.
  - Evaluation was done for individual segments.

H2020 RIA ELITR (2019–2021)

EU research and innovation project ELITR aims:

- Highly multi-lingual machine and speech translation.
- Document-level machine translation.
- Automatic meeting summarization, “Minuting”.

Partners:

- CUNI: Coordinator, MT, Minuting.
- UEDIN: MT, Spoken Language Translation.
- KIT: ASR, Spoken Language Translation.
- Pervoice: Integrator, platform for channeling speech.
- alfatraining: User, Remote conferencing platform.
- Czech Supreme Audit Office: User, Face-to-face and remote conferences.

ELITR Languages

Languages to be supported by the ELITR project:

24 EU languages: Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovene, Spanish, and Swedish.
Further 19 EUROSAI languages: Albanian, Arabic, Armenian, Azerbaijani, Belorussian, Bosnian, Georgian, Hebrew, Icelandic, Kazakh, Luxembourgish, Macedonian, Moldovan, Montenegrin, Norwegian, Russian, Serbian, Turkish, and Ukrainian.

Technology	Primary Focus	Covered	Experimental
ASR	En, De	Fr, Sp, It, Ru	Cs
SLT & MT	{ En, De } $\rightarrow$ { En, De, Cs }	all EU languages $\rightarrow$ all EU languages	all EUROSAI langs $\rightarrow$ all EUROSAI langs
Summarization	English, Czech	–	–

Some of ELITR Speech Translation Challenges

Input acquisition:

- Soud quality, speaker specifics, dialects, ...
- ASR quality:
  - Training data, training data, training data, ...
  - Jargon, abbreviations, ...

Spoken Language

- Segmentation, adding punctuation.
- Latency:
  - A trade-off: translate sooner with less context or better wait?
- All the problems of MT: Ambiguity, context, rare words, ...

Delivery:

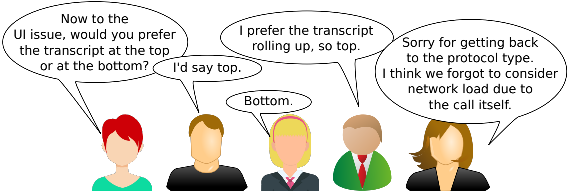
- Wifi failing in crowded areas.
- People cannot follow slides, speaker and subtitles on handheld device.

Minuting

- Input: A hierarchical agenda.
- Speech recorded of the fly.
  - Distinguishing speakers.
- Full transcript available
  - Possibly for immediate manual correction.
- Minuting: Agenda gets populated with summaries of statements.

Difference from summarization:

- We want to keep all information, only deduplicate.



Original agenda as prepared by the organizer beforehand:
- Protocol type: push or pull?
- Layout of the user interface:
  - Transcript grows at the top or bottom of the document?
  - Or in a side pane?

Shared document, everyone allowed to edit.
Starts with the agenda and gets populated by Automatic Minuting (AM)
- Protocol type: push or pull?
  - AM > Pull easier to implement.
  - AM > Updates can get lost with push *in case the user*
  - AM > Consider network load.
- Layout of the user interface:
  - Transcript grows at the top or bottom of the document?
    - AM > Top AM > Bottom AM > Top, transcript rolling up.
  - Or in a side pane?

Transcript, optionally editable to correct ASR errors:
11:03 Sorry for putting back to the protocol type. I think we forgot ...
11:02 I prefer the transcript rolling up, so top.
11:02 Bottom
...

Summarization as Machine Translation

Input:

legendární slovenská punkrocková kapela extip se letos vrátila na pódia poté, co vyšla v reedici její debutová deska pekný, škaredý deň, kterou přehraje 1. prosince na sedmičce na strahově. soubor nezanikl, i když bratislavskou punkovou scénu v devadesátých letech rozložily drogy. své zkušenosti s tím má kytarista sveto korbel, který odpovídal na otázky novinek.

Human Output:

slovenská punková legenda extip se vrátila

Summarization as Machine Translation

Input:
legendární slovenská punkrocková kapela extip se letos vrátila na pódia poté, co vyšla v reedici její debutová deska pekný, škaredý deň, kterou přehraje 1. prosince na sedmičce na strahově. soubor nezanikl, i když bratislavskou punkovou scénu v devadesátých letech rozložily drogy. své zkušenosti s tím má kytarista sveto korbel, který odpovídal na otázky novinek.
Human Output:
slovenská punková legenda extip se vrátila
“Summarized” by MT from paragraphs to sentences:
slovenská kapela extip se vrací do prahy

Meaning Understood?

Input:
legendární slovenská punkrocková kapela extip se letos vrátila na pódia poté, co vyšla v reedici její debutová deska pekný, škaredý deň, kterou přehraje 1. prosince na sedmičce na **strahově**. soubor nezanikl, i když bratislavskou punkovou scénu v devadesátých letech rozložily drogy. své zkušenosti s tím má kytarista sveto korbel, který odpovídal na otázky novinek.

Human Output:
slovenská punková legenda extip se vrátila

“Summarized” by MT from paragraphs to sentences:
slovenská kapela extip se vrací do **prahy**

Not Understood.

na strahově	slovenská kapela extip se vrací do **prahy**
v o2 aréně	slovenská kapela extip se vrací do **prahy**
na hradecku	slovenská kapela extip se vrací do **čech**
u vajgaru	slovenská kapela extip se vrací do **prahy**

Not Understood.

na strahově	slovenská kapela extip se vrací do **prahy**
v o2 aréně	slovenská kapela extip se vrací do **prahy**
na hradecku	slovenská kapela extip se vrací do **čech**
u vajgaru	slovenská kapela extip se vrací do **prahy**
ve stromovce	

slovenská kapela extip se vrací na scénu. tentokrát kvůli drogám v reedici. s. s. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. i. m. . . . . m. . . . m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. . m. m. m. m. m. m. m. m. m. m.

Summary

- Deep Learning allowed tremendous jump in quality.
  - Thanks to computing power and available data.
- In principle, it can "learn representations".
  - … but it excels at overfitting.
  - So far, it is not learning to understand, **by no means at all**.
- Yes, humans can be surpassed by machines.
  - In NMT, this has been reached in very specific circumstances.
  - The test setting has to closely match the training setting.
  - The systems can spectacularly fall apart if outside of their "comfort zone".

ELITR needs you:

- Experience, real usecases, opportunities for testing (off-line or live).
- Data:
  - Speech recordings and transcripts.
  - Meeting recordings, transcripts, agendas, minutes.

References

Ondřej Cífka and Ondřej Bojar. 2018. Are BLEU and Meaning Representation in Opposition? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1362–1371. Association for Computational Linguistics, Association for Computational Linguistics.

Bonnie j. Dorr, Rebecca j. Passonneau, David Farwell, Rebecca Green, Nizar Habash, Stephen Helmreich, Eduard Hovy, Lori Levin, Keith j. Miller, Teruko Mitamura, Owen Rambow, and Advaith Siddharthan. 2010. Interlingual annotation of parallel text corpora: A new framework for annotation and evaluation. *Nat. Lang. Eng.*, 16(3):197–243.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR*, abs/1611.04558.

Tom Kocmi and Ondřej Bojar. 2017. Curriculum Learning and Minibatch Bucketing in Neural Machine Translation. In *Proceedings of Recent Advances in NLP (RANLP 2017)*.

Andrew Lampert. 2001. Interlingua in Machine Translation. http://sgi.nu/nlp/content/pdf/InterlinguaInMachineTranslation.pdf, September.

Jan Niehues, Eunah Cho, Thanh-Le Ha, and Alex Waibel. 2016. Pre-translation for neural machine translation. *CoRR*, abs/1610.05243.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

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