Building practical MT systems for real life: research and/or engineering?

Oravecz Csaba
Westpole Luxembourg
oravecz.csaba@gmail.com

MILAB-NLP seminar

Zoom, 25 March 2021

Outline

1. Introduction

2. Systems for real life
    - Workflows for building practical systems
    - High quality (impractical) NMT systems

3. Methods to improve translation quality, system efficiency and service
    - Data centric approaches
    - Model centric approaches
    - Auxiliary tasks
    - Future directions

4. Questions with(out) answers

Motivation

Overview of how (im)practical (high quality) NLP systems are produced

- basic (or advanced) workflows in practical NMT systems
- issues, errors, practical solutions
- HOWTO: win the WMT news shared task

Questions (with or without answers)

- What are the most important current research problems in NMT?
- How does SOTA research results carry over to (our) real life systems?
- Do we need something special to deal with Hungarian?
- Does linguistics have a place in practical MT systems?

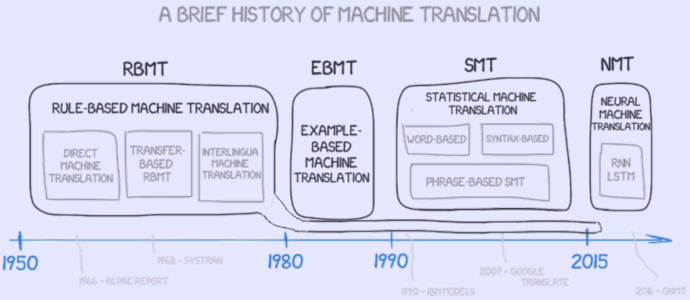
Background

Project

- CEF eTranslation project
- MT service between all 26 official languages of the EU and the EEA for translators and officials in EU and national authorities
- EU formal and general language engines
- recently extended to Zh, Ja, Ar, Tr ($\sum > 100$ systems)
- domain specific engines (health, law, financial)
- runs in MS Azure

Background

The briefest history of all[1]



[1] Source: `http://vas3k.com/blog/machine_translation/`

MT evolution

Translate:

MT evolution

Translate:

It is a prerequisite for enrolment in the training that there is in place a practical training agreement between the participant and a regional body of the Danish Veterinary and Food Administration (Fødevarestyrelsen).

MT evolution

Translate:

It is a prerequisite for enrolment in the training that there is in place a practical training agreement between the participant and a regional body of the Danish Veterinary and Food Administration (Fødevarestyrelsen).

Rule based

Ez egy előfeltétel beiratkozásért bent van az edzeni azt ott helyben van egy gyakorlati tréningmegegyezés a résztvevő és a dán Veterinary és Food Administration (Fødevarestyrelsen). egy területi teste között

MT evolution

Translate:

It is a prerequisite for enrolment in the training that there is in place a practical training agreement between the participant and a regional body of the Danish Veterinary and Food Administration (Fødevarestyrelsen).

PBMT

A nyilvántartásba vétel előfeltétele, hogy a képzés olyan gyakorlati képzésben résztvevő közötti megállapodás egy regionális szerv és a dán állategészségügyi és élelmiszerügyi hatóságot (Fødevarestyrelsen).

MT evolution

Translate:

It is a prerequisite for enrolment in the training that there is in place a practical training agreement between the participant and a regional body of the Danish Veterinary and Food Administration (Fødevarestyrelsen).

NMT yesterday

A képzésen való részvétel előfeltétele, hogy gyakorlati képzési megállapodás jöjjön létre a résztvevő és a Dán Állategészségügyi és Élelmezésügyi Hivatal regionális szerve között (Fødevarestyrelsen).

MT evolution

Translate:

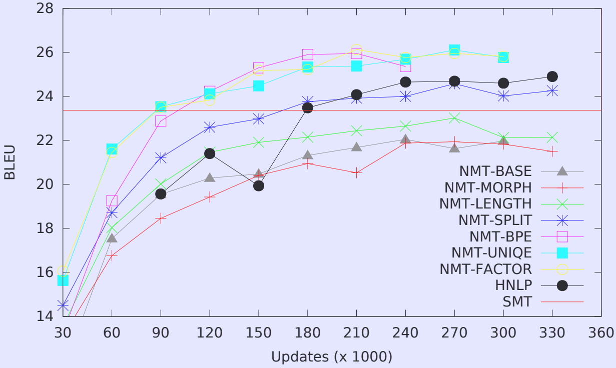
It is a prerequisite for enrolment in the training that there is in place a practical training agreement between the participant and a regional body of the Danish Veterinary and Food Administration (Fødevarestyrelsen).

NMT today

A képzésben való részvétel előfeltétele, hogy a résztvevő és a Dán Állat-egészségügyi és Élelmiszerügyi Hivatal (Fødevarestyrelsen) regionális szerve között gyakorlati képzési megállapodás jöjjön létre.

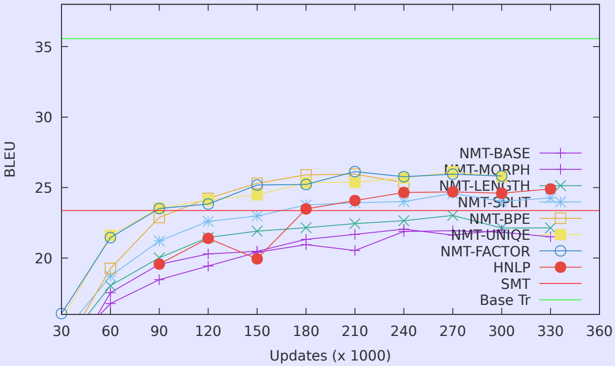
Score evolution

MSZNY 2017



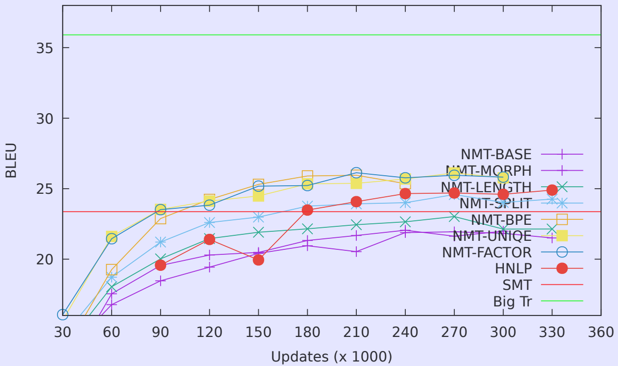
Score evolution

Base transformer



Score evolution

Big transformer



Outline

Data

EU formal: high quality parallel data

- EURAMIS database [Steinberger et al., 2012]
  - manual translations from EU institutions
  - 3m (Ga) – 40m (Fr) segments (Hu: 22m)

General: mixed quality parallel data

- data from all over the place (mostly OPUS)
- ParaCrawl [Esplà et al., 2019, Bañón et al., 2020]
- 1m (Ga) – >100m (Fr, De, Es)

Data filtering

Monolingual cleanup — cheap tricks

- minimum number of alphabetic characters, unicode filter
- maxlength (60-150 tokens→subwords)
- character/token ratio ([1.5,40])
- language identification (fasttext)

Parallel data scoring and filtering

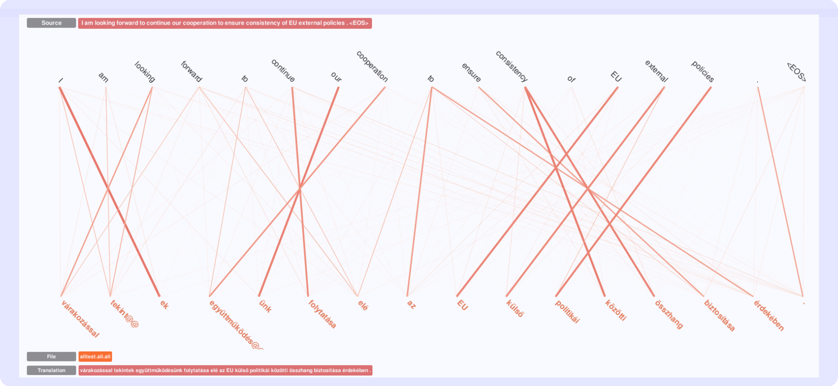
- standard tool: Bicleaner [Esplà-Gomis et al., 2020]
  rule-based prefilter, LM based fluency scorer, random forest classifier
- similarity scoring based on sentence embeddings [Zhang et al., 2020, Guo et al., 2018]

Data pre- and postprocessing

EU formal systems: $\lim_{n\to\infty} U(n) = 1$

- n: number of PP steps; U: user satisfaction ([0,1])
- standard steps: tokenization, normalization, truecasing
- placeholders (?)
  - masking of specific patterns (losing semantic content and context)
  - (soft) alignment based target side replacement (use the Hungarian (Munkres) algorithm!)

Attention, attention



Data pre- and postprocessing

EU formal systems: $\lim_{n \to \infty} U(n) = 1$

- n: number of PP steps; U: user satisfaction ([0,1])
- standard steps: tokenization, normalization, truecasing
- placeholders (?)
  - masking of specific patterns (losing semantic content and context)
  - (soft) alignment based target side replacement (use the Hungarian (Munkres) algorithm!)

Otherwise

- minimal or $\emptyset$

Vocabulary

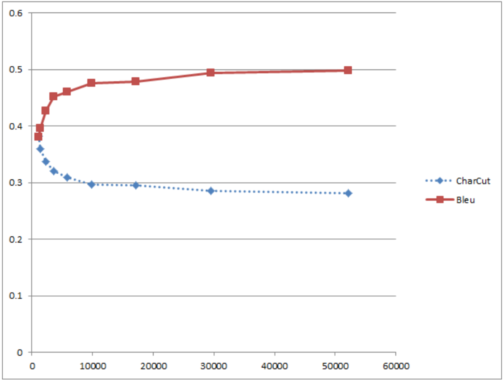
Rare words are very common $\rightarrow$ subword segmentation

- optimal size? [Gowda and May, 2020]
- BPE: greedy segmentation [Sennrich et al., 2016b] (on tokenized input)
- SentencePiece: unigram LM based segmentation [Kudo and Richardson, 2018] (on raw input)
- Google's WordPiece (not widely used in MT)

Subword regularization

- utilize the segmentation ambiguity as a noise to improve the robustness of NMT [Kudo, 2018]
- BPE-Dropout [Provilkov et al., 2020]: stochastically corrupts BPE segmentation $\rightarrow$
  multiple segmentations within the same fixed BPE framework

Vocabulary sizes and scores



Training and inference

Basic setup

- base transformer with standard hyperparameter settings
- 32–36k BPE/SP joint vocabulary
- 2–4 V100 GPUs
- train until sentence-wise normalized cross-entropy stalls on the validation set for 5(–10) validation steps

Decoding speed

- low resource environment, implementation constraints
- default 6 layer self-attention decoder too slow
- reduced layer number for morphorich languages, RNN decoder for others
- somewhat reduced quality, 2–4 x speedup

Evaluation

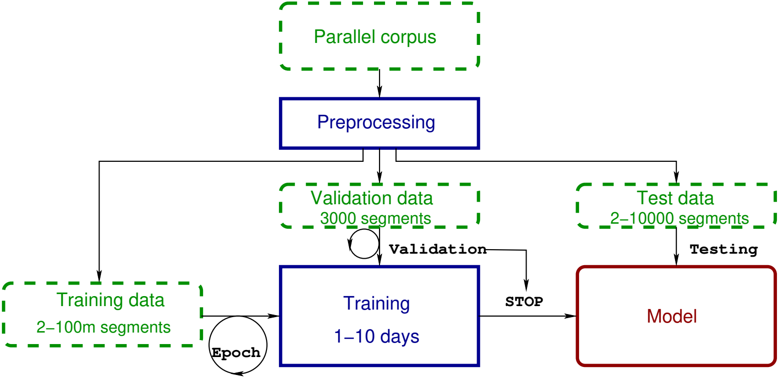
Cross-evaluation on fixed test sets

- BLEU scores:
  - EU formal engines: 45–70 (worst: Fi, Hu; best: Mt, Pt)
  - generic: 30–50 (worst: Fi, Hu; best: Mt, Es)

Issues, frequent errors

- fluency vs adequacy
- domain robustness
- long segments: under- or overtranslation, hallucination $\rightarrow$ automatic segmentation? [Pouget-Abadie et al., 2014] big bird transformer? [Zaheer et al., 2020]
- named entities (placeholders?)
- input formatting/markup (brrr...) [Hanneman and Dinu, 2020]

The Engine Factory



Outline

High quality NMT system HOWTO

How to win the WMT News Task

- high peformance WMT engines
  - some (standard) data filtering
  - (iterative) (tagged) back-translation (forward translation does not work)
  - (ensembles of) huge (very deep) models (up to 50 encoder layers, 15000 FFN, 256 heads)
    (cf. base: 6 layers, 2048 FFN, 8 heads)
  - domain fine tuning

Data selection vs. model complexity

Domain specific data rulez (even if noisy)

- build Fr–De system for European election news
- tune base model towards the topic by making use of guided topic modeling [Jagarlamudi et al., 2012]
  - seed word list from German news articles on elections
  - classifiy documents in the 2014 and 2016 German News Crawl into topics
  - select candidate data (4m segments) for back-translation
- significant increase on task test set (3.7 BLEU)

Result

French→German		
Ave.	Ave. z	System
82.4	0.267	MSRA-MADL
81.5	0.246	eTranslation
78.5	0.082	LIUM
76.8	0.037	MLLP-UPV
76.0	0.001	online-Y
76.6	−0.018	online-G
75.2	−0.034	online-B
74.8	−0.039	online-A
73.9	−0.098	TartuNLP-c
66.5	−0.410	online-X

All in

En–De: the highest resource system

- 46m OP, 500m in domain mono, 35k dev set
- (normally) the strongest competition ($\approx$ Zh)
- stepwise development from simple to complex models [Oravecz et al., 2020]
  - base transformer from OP
  - + (tagged) back-translation
  - + continued training on LM scored and ranked OP subset until BLEU increases
  - + big transformer
  - + fine tuning
  - + ensembling

Evolution of models

En–De results

System	Data	Test sets 2019	Test sets 2020
M1: Baseline	44.7M	41.9	32.7
M2: M1+BT+CT	64.7M	43.3	34.4
M3: M2+Tbig	232M	44.5	36.9
M4: M3+FT	232M+34.5k	44.8	37.2
M5: M4 ens	232M+34.5k	**46.0**	**37.9**→(38.8)

Evolution of models

En–De results

		Test sets	
System	Data	2019	2020
M1: Baseline	44.7M	41.9	32.7
M2: M1+BT+CT	64.7M	43.3	34.4
M3: M2+Tbig	232M	44.5	36.9
M4: M3+FT	232M+34.5k	44.8	37.2
M5: M4 ens	232M+34.5k	**46.0**	**37.9**$\rightarrow$(38.8)

base tr. with OP

Evolution of models

En–De results

base tr. with back translation and continued training

System	Data	Test sets 2019	Test sets 2020
M1: Baseline	44.7M	41.9	32.7
M2: M1+BT+CT	64.7M	43.3	34.4
M3: M2+Tbig	232M	44.5	36.9
M4: M3+FT	232M+34.5k	44.8	37.2
M5: M4 ens	232M+34.5k	**46.0**	**37.9**→(38.8)

Evolution of models

En–De results

System	Data	Test sets 2019	2020
M1: Baseline	44.7M	41.9	32.7
M2: M1+BT+CT	64.7M	43.3	34.4
M3: M2+Tbig	232M	44.5	36.9
M4: M3+FT	232M+34.5k	44.8	37.2
M5: M4 ens	232M+34.5k	**46.0**	**37.9**→(38.8)

big tr. with more back translation and continued training

Evolution of models

En–De results

System	Data	Test sets 2019	Test sets 2020
M1: Baseline	44.7M	41.9	32.7
M2: M1+BT+CT	64.7M	43.3	34.4
M3: M2+Tbig	232M	44.5	36.9
M4: M3+FT	232M+34.5k	44.8	37.2
M5: M4 ens	232M+34.5k	**46.0**	**37.9**→(38.8)

big tr. with more back translation, continued training and fine tuning

Evolution of models

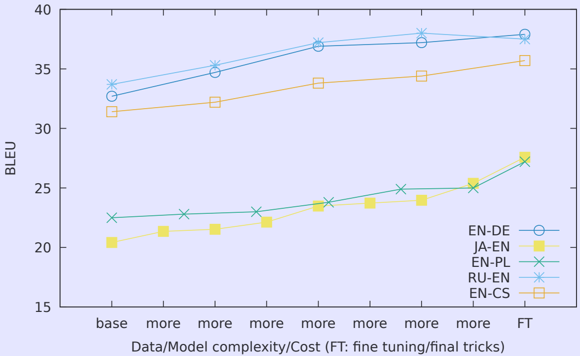
En–De results

		Test sets	
System	Data	2019	2020
M1: Baseline	44.7M	41.9	32.7
M2: M1+BT+CT	64.7M	43.3	34.4
M3: M2+Tbig	232M	44.5	36.9
M4: M3+FT	232M+34.5k	44.8	37.2
M5: M4 ens	232M+34.5k	**46.0**	**37.9**→(38.8)

ensemble of best individual models

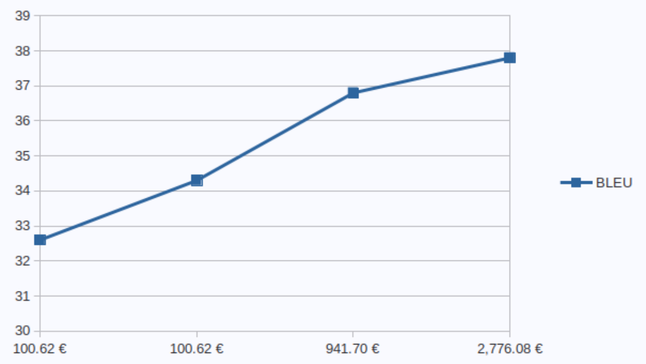
Evolution of models

All in one



TINSTAAFL

The No Free Lunch Theorem for Machine Translation



WMT lessons

- cheaper to focus on data than on “smart” models
- no established best practice to rule them all
- differences between systems are small, cannot control for all parameter settings (including data related processing) $\rightarrow$ accidentally finding some optimal (best test set fitting) configuration
- customized solutions most importantly wrt data selection and filtering
- top systems are rarely suitable for large scale production

Outline

Trends


DeepLearning.AI
A CHAT WITH ANDREW
MLOps: From Model-centric to Data-centric AI
Wed, MARCH 24
10 to 11am PT
RSVP: mlops0324.eventbrite.com
ANDREW NG
Founder
DeepLearning.AI

Clean data — better engine

Data quality

- NMT sensitive to noise [Khayrallah and Koehn, 2018, Koehn and Knowles, 2017]
- improving the quality of training data by removing spurious translations
- data filtering from noisy parallel data $\rightarrow$ separate WMT shared task

More data — better engine

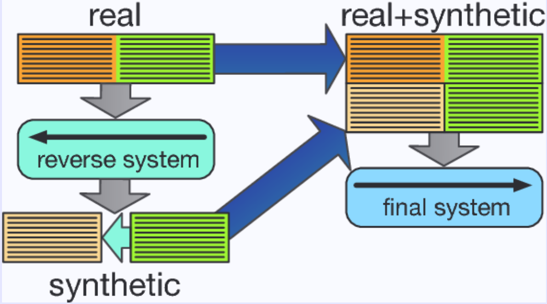
Back-translation (BT)

- improving NMT with monolingual data [Sennrich et al., 2016a]
- tagged back translation [Caswell et al., 2019, Marie et al., 2020]
  BT introduces:
  - helpful signal (strong target-language, weak cross-lingual signal)
  - harmful signal (amplifying MT bias)

  BT label allows the model to separate helpful and harmful signal
- iterative BT [Hoang et al., 2018]

BT

Simple but effective[2]



[2] Source: Hoang et al. [2018]

More data — better engine

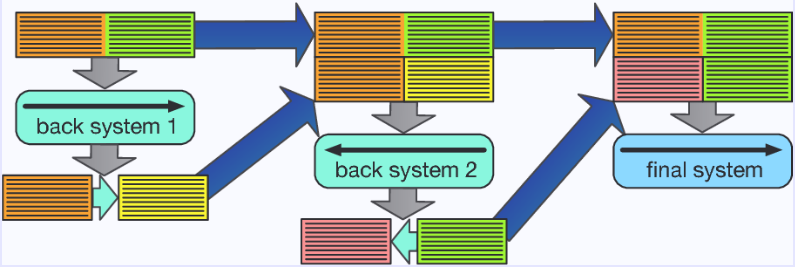
Back-translation (BT)

- improving NMT with monolingual data [Sennrich et al., 2016a]
- tagged back translation [Caswell et al., 2019, Marie et al., 2020]
  BT introduces:
  - helpful signal (strong target-language, weak cross-lingual signal)
  - harmful signal (amplifying MT bias)

  BT label allows the model to separate helpful and harmful signal
- iterative BT [Hoang et al., 2018]

Iterative BT

Complex but can be effective[3]



[3] Source: Hoang et al. [2018]

Task specific data — better engine

Domain adaptation (transfer learning)

- domain robustness [Müller et al., 2020]
  - SMT: mostly adequate but not fluent
    NMT: mostly fluent, but not adequate
  - hallucinations (translations that are fluent but unrelated to the source): key reason for low domain robustness
  - various methods with mixed results
    - subword regularization, defensive distillation, reconstruction, n-best list reranking
  - “radically different approaches are needed to increase the coverage and adequacy of NMT translations without sacrificing their fluency”

in-domain and *out-domain* translation

Translate in the health domain

in-domain and *out-domain* translation

Translate in the health domain

An example of its use is the treatment of some type of tumours, where the radiolabelled medicine carries the radioactivity to the site of a tumour to destroy the tumour cells.

in-domain and *out-domain* translation

Translate in the health domain

An example of its use is the treatment of some type of tumours, where the radiolabelled medicine carries the radioactivity to the site of a tumour to destroy the tumour cells.

Out of domain NMT yesterday

Ennek egyik példája a daganattípusok kezelése, ahol a radioizotóppal jelölt gyógyszer a tumor helyszínének radioaktivitását hordozza, hogy elpusztítsa a tumor sejteket.

in-domain and *out-domain* translation

Translate in the health domain

An example of its use is the treatment of some type of tumours, where the radiolabelled medicine carries the radioactivity to the site of a tumour to destroy the tumour cells.

Out of domain NMT today

Egy példa a daganatok bizonyos típusainak kezelésére, ahol a radioizotóppal jelölt gyógyszer radioaktivitást visz a daganat helyszínére a daganatsejtek elpusztítása céljából.

in-domain and *out-domain* translation

Translate in the health domain

An example of its use is the treatment of some type of tumours, where the radiolabelled medicine carries the radioactivity to the site of a tumour to destroy the tumour cells.

In-domain NMT (yesterday)

Egyik alkalmazása lehet bizonyos fajta daganatok kezelése, ahol a radioaktív izotóppal jelzett gyógyszer a daganat területére szállítja a radioaktivitást.

Previously...

Translate:

It is a prerequisite for enrolment in the training that there is in place a practical training agreement between the participant and a regional body of the Danish Veterinary and Food Administration (Fødevarestyrelsen).

Previously...

Translate:

It is a prerequisite for enrolment in the training that there is in place a practical training agreement between the participant and a regional body of the Danish Veterinary and Food Administration (Fødevarestyrelsen).

General NMT

A képzésben való részvétel előfeltétele, hogy a résztvevő és a Dán Állat-egészségügyi és Élelmiszerügyi Hivatal (Fødevarestyrelsen) regionális szerve között gyakorlati képzési megállapodás jöjjön létre.

Previously...

Translate:

It is a prerequisite for enrolment in the training that there is in place a practical training agreement between the participant and a regional body of the Danish Veterinary and Food Administration (Fødevarestyrelsen).

Health domain NMT

Előfeltétele annak, hogy a felállás egy olyan gyakorlati képzés formájában hozták létre, amelyet a dán állatgyógyászati és Élelmiszer-biztonsági hatóság regionális, valamint a dán alkalmazások és az élelmiszer- és élelmiszerbiztonsági hatóság közötti gyakorlati képzés felel meg (amely

Task specific data — better engine

Domain adaptation (transfer learning)

- catastrophic forgetting [Thompson et al., 2019, Gu and Feng, 2020]
- domain adaptation with residual adapters; small adapter components that are plugged in each hidden layer $\rightarrow$ adapters are trained only with the in-domain data, keeping the pretrained model frozen
  [Bapna and Firat, 2019, Pham et al., 2020a]

Outline

Methods

Increasing model complexity

- deeper models (mostly encoder) to model more complex dependencies; vanishing gradient [Zhang et al., 2019, Liu et al., 2020a]
- layer normalization [Ba et al., 2016]: Post-LN and Pre-LN Transformer

Increasing training complexity

- teacher student training, knowledge distillation [Freitag et al., 2017]
  - sequence level KD: student trained on teacher output with highest score
  - sequence level interpolation: student trained on teacher output most similar to gold target
- multilingual [Lepikhin et al., 2020], multidomain models (transfer learning)
  multidimensional tagging [Stergiadis et al., 2021]

Training with the wrong objective?

Standard training objective

- minimize the negative log-likelihood $\mathscr{L}(\theta)$ of the training data D

$$\mathscr{L}(\theta) = \sum_{(\mathbf{x},\mathbf{y}) \in D} \sum_{t=1}^{|\mathbf{y}|} -\log P(\mathbf{y}_t | \mathbf{x}, \mathbf{y}_{<t}; \theta)$$

$\mathbf{x}, \mathbf{y}$: source and target sequence
$\mathbf{y}_t$: t$^{\text{th}}$ token in $\mathbf{y}$
$\mathbf{y}_{<t}$: all previous tokens
MLE with teacher forcing: $\mathbf{y}_{<t}$ ground-truth labels in training $\rightarrow$ mismatch with inference ($\mathbf{y}_{<t}$ model predictions)
$\rightarrow$ exposure bias [Wang and Sennrich, 2020]

Minimum Risk Training

This is the way.

- MRT: sequence level objective, objective function: expected loss (*risk*) wrt posterior distribution [Shen et al., 2016]

$$\mathscr{R}(\theta) = \sum_{(\mathbf{x},\mathbf{y}) \in D} \sum_{\tilde{\mathbf{y}} \in \mathscr{Y}(\mathbf{x})} P(\tilde{\mathbf{y}}|\mathbf{x}) \Delta(\tilde{\mathbf{y}}, \mathbf{y})$$

$\Delta(\tilde{\mathbf{y}}, \mathbf{y})$: discrepancy between gold translation $\mathbf{y}$ and model prediction $\tilde{\mathbf{y}}$

Minimum Risk Training

Training objective

- minimize the risk on the training data
- search space is intractable $\rightarrow$ posterior distribution $\mathscr{Y}(\mathbf{x})$ is approximated by a subspace by sampling a certain number of candidate translations and normalizing
- loss: $1-$*sentence level smoothed BLEU*
- Marian: "MRT and Reinforcement Learning are things I always want to do, but never have time to implement."

Other issues

- translation efficiency [Kim et al., 2019]
- data noising [Xie et al., 2017]
- terminology support
  - training time [Dinu et al., 2019, Exel et al., 2020, Bergmanis and Pinnis, 2021]
  - decoding time: constrained decoding [Post and Vilar, 2018, Hokamp and Liu, 2017]

Outline

Evaluating and improving MT output

- QA (Quality Estimation)
  - estimate MT output quality without reference
  - word level (+/-); segment level (score)
- APE (Automatic Post-Editing)

Outline

Remaining problems

- long range dependencies
- long segments
- terminology (adequacy)
- unsupervised MT [Marchisio et al., 2020]
- zero-shot MT
- multimodal (simultaneous) translation [Imankulova et al., 2020]
- more efficient transformer architectures [Tay et al., 2020]

Most promising developments

- document level MT [Lopes et al., 2020, Ma et al., 2021]
- discourse level MT [Zhang, 2020]
- document/instance based dynamic domain adaptation [Farajian et al., 2017, Xu et al., 2019, Pham et al., 2020b]
- adaptive MT
  - live model training during post-editing from $<$MT output, PE output$>$ pairs
  - expensive (time and resource)

Qs

- What are the most important current research problems in NMT?
- How does SOTA research results carry over to (our) real life systems?
  - large focus on low resource settings in research
  - many results do not carry over to practically usable models (which are best case trained with substantial data or high quality parallel domain data)
- Do we need something special to deal with Hungarian?
- Does linguistics have a place in practical MT systems?
  "Phonemes are a fantasy of linguists" (Andrew Ng)
- Pretrained LMs in NMT [Liu et al., 2020b]
  seem to work only in low resource or (some) multilingual settings; the more data the less gain (or even drop) in quality

Papers with code...

README.md

Code/models are under review by the MS legal team now, and they will be released once it is done.

Qs

- What are the most important current research problems in NMT?
- How does SOTA research results carry over to (our) real life systems?
  - large focus on low resource settings in research
  - many results do not carry over to practically usable models (which are best case trained with substantial data or high quality parallel domain data)
- Do we need something special to deal with Hungarian?
- Does linguistics have a place in practical MT systems?
  "Phonemes are a fantasy of linguists" (Andrew Ng)
- Pretrained LMs in NMT [Liu et al., 2020b]
  seem to work only in low resource or (some) multilingual settings; the more data the less gain (or even drop) in quality

Confusions and future steps

Conclusions and future steps

Conclusions and future steps

Lessons learned

- reasonable models can be produced using established techniques even in very constrained conditions
- brute force is useful but expensive
- data selection is more rewarding and a lot cheaper

Conclusions and future steps

Lessons learned

- reasonable models can be produced using established techniques even in very constrained conditions
- brute force is useful but expensive
- data selection is more rewarding and a lot cheaper

Where we are going we don't need roads

- magical data generation algorithm, which out of an empty set, generates a high quality parallel data set of any amount for any language

I have spoken.

References I

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.417. URL `https://www.aclweb.org/anthology/2020.acl-main.417`.

References II

Ankur Bapna and Orhan Firat. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1165. URL https://www.aclweb.org/anthology/D19-1165.

Toms Bergmanis and Mārcis Pinnis. Facilitating terminology translation with target lemma annotations, 2021.

Isaac Caswell, Ciprian Chelba, and David Grangier. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5206. URL https://www.aclweb.org/anthology/W19-5206.

References III

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1294. URL `https://www.aclweb.org/anthology/P19-1294`.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland, August 2019. European Association for Machine Translation. URL `https://www.aclweb.org/anthology/W19-6721`.

References IV

Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena, Jaume Zaragoza-Bernabeu, and Felipe Sánchez-Martínez. Bicleaner at WMT 2020: Universitat d'alacant-prompsit's submission to the parallel corpus filtering shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 952–958, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.wmt-1.107.

Miriam Exel, Bianka Buschbeck, Lauritz Brandt, and Simona Doneva. Terminology-constrained neural machine translation at SAP. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 271–280, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://www.aclweb.org/anthology/2020.eamt-1.29.

References V

M. Amin Farajian, Marco Turchi, Matteo Negri, and Marcello Federico. Multi-domain neural machine translation through unsupervised adaptation. In *Proceedings of the Second Conference on Machine Translation*, pages 127–137, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4713. URL https://www.aclweb.org/anthology/W17-4713.

Markus Freitag, Yaser Al-Onaizan, and Baskaran Sankaran. Ensemble distillation for neural machine translation. *CoRR*, abs/1702.01802, 2017. URL http://arxiv.org/abs/1702.01802.

Thamme Gowda and Jonathan May. Finding the optimal vocabulary size for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3955–3964, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.352. URL https://www.aclweb.org/anthology/2020.findings-emnlp.352.

References VI

Shuhao Gu and Yang Feng. Investigating catastrophic forgetting during continual training for neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4315–4326, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.381. URL `https://www.aclweb.org/anthology/2020.coling-main.381`.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6317. URL `https://www.aclweb.org/anthology/W18-6317`.

References VII

Greg Hanneman and Georgiana Dinu. How should markup tags be translated? In *Proceedings of the Fifth Conference on Machine Translation*, pages 1160–1173, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.wmt-1.138.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2703. URL https://www.aclweb.org/anthology/W18-2703.

Chris Hokamp and Qun Liu. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1141. URL https://www.aclweb.org/anthology/P17-1141.

References VIII

Aizhan Imankulova, Masahiro Kaneko, Tosho Hirasawa, and Mamoru Komachi. Towards multimodal simultaneous neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 594–603, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.wmt-1.70.

Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213, Avignon, France, April 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E12-1021.

References IX

Huda Khayrallah and Philipp Koehn. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2709. URL https://www.aclweb.org/anthology/W18-2709.

Young Jin Kim, Marcin Junczys-Dowmunt, Hany Hassan, Alham Fikri Aji, Kenneth Heafield, Roman Grundkiewicz, and Nikolay Bogoychev. From research to production and back: Ludicrously fast neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 280–288, Hong Kong, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5632. URL https://www.aclweb.org/anthology/D19-5632.

References X

Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3204. URL `https://www.aclweb.org/anthology/W17-3204`.

Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1007. URL `https://www.aclweb.org/anthology/P18-1007`.

References XI

Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL `https://www.aclweb.org/anthology/D18-2012`.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding, 2020.

Xiaodong Liu, Kevin Duh, Liyuan Liu, and Jianfeng Gao. Very deep transformers for neural machine translation, 2020a.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation, 2020b.

References XII

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL `https://www.aclweb.org/anthology/2020.eamt-1.24`.

Zhiyi Ma, Sergey Edunov, and Michael Auli. A comparison of approaches to document-level machine translation, 2021. URL `https://arxiv.org/abs/2101.11040`.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583, Online, November 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.wmt-1.68`.

References XIII

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. Tagged back-translation revisited: Why does it really work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5990–5997, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.532. URL https://www.aclweb.org/anthology/2020.acl-main.532.

Mathias Müller, Annette Rios, and Rico Sennrich. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 151–164, Virtual, October 2020. Association for Machine Translation in the Americas. URL https://www.aclweb.org/anthology/2020.amta-research.14.

References XIV

Csaba Oravecz, Katina Bontcheva, László Tihanyi, David Kolovratnik, Bhavani Bhaskar, Adrien Lardilleux, Szymon Klocek, and Andreas Eisele. eTranslation's submissions to the WMT 2020 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 254–261, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.wmt-1.26.

Minh Quang Pham, Josep Maria Crego, François Yvon, and Jean Senellart. A study of residual adapters for multi-domain neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 617–628, Online, November 2020a. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.wmt-1.72.

References XV

Minh Quang Pham, Jitao Xu, Josep Crego, François Yvon, and Jean Senellart. Priming neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 516–527, Online, November 2020b. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.wmt-1.63`.

Matt Post and David Vilar. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1119. URL `https://www.aclweb.org/anthology/N18-1119`.

References XVI

Jean Pouget-Abadie, Dzmitry Bahdanau, Bart van Merriënboer, Kyunghyun Cho, and Yoshua Bengio. Overcoming the curse of sentence length for neural machine translation using automatic segmentation. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 78–85, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4009. URL https://www.aclweb.org/anthology/W14-4009.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.170. URL https://www.aclweb.org/anthology/2020.acl-main.170.

References XVII

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August 2016a. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://www.aclweb.org/anthology/P16-1009.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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 2016b. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://www.aclweb.org/anthology/P16-1162.

References XVIII

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1159. URL https://www.aclweb.org/anthology/P16-1159.

Ralf Steinberger, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlüter. DGT-TM: A freely available translation memory in 22 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 454–459, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/814_Paper.pdf.

Emmanouil Stergiadis, Satendra Kumar, Fedor Kovalev, and Pavel Levin. Multi-domain adaptation in neural machine translation through multidimensional tagging, 2021.

References XIX

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey, 2020.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2062–2068, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1209. URL https://www.aclweb.org/anthology/N19-1209.

Chaojun Wang and Rico Sennrich. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.326. URL https://www.aclweb.org/anthology/2020.acl-main.326.

References XX

Ziang Xie, Sida I. Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y. Ng. Data noising as smoothing in neural network language models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=H1VyHY9gg`.

Jitao Xu, Josep Crego, and Jean Senellart. Lexical Micro-adaptation for Neural Machine Translation. In *International Workshop on Spoken Language Translation*, International Workshop on Spoken Language Translation, Hong Kong, Hong Kong SAR China, November 2019. URL `https://hal.archives-ouvertes.fr/hal-02635039`.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33, 2020.

References XXI

Biao Zhang, Ivan Titov, and Rico Sennrich. Improving deep transformer with depth-scaled initialization and merged attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 898–909, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1083. URL https://www.aclweb.org/anthology/D19-1083.

Boliang Zhang, Ajay Nagesh, and Kevin Knight. Parallel corpus filtering via pre-trained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8545–8554, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.756. URL https://www.aclweb.org/anthology/2020.acl-main.756.

References XXII

Xiaojun Zhang. A review of discourse-level machine translation. In *Proceedings of the Second International Workshop of Discourse Processing*, pages 4–12, Suzhou, China, December 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.iwdp-1.2.